MTH 581-582
Introduction to Abstract Algebra

D. S. Malik
Creighton University

John N. Mordeson
Creighton University

M.K. Sen
Calcutta University

Department of Mathematics

Printed in the United States of America

This document was produced with *Scientific Word*.

Preface

This book is intended for a one-year introductory course in abstract algebra with some topics of an advanced level. Its design is such that the book can also be used for a one-semester course. The book contains more material than normally would be taught in a one-year course. This should give the teacher flexibility with respect to the selection of the content and the level at which the book is to be used. We give a rigorous treatment of the fundamentals of abstract algebra with numerous examples to illustrate the concepts. It usually takes students some time to become comfortable with the seeming abstractness of modern algebra. Hence we begin at a leisurely pace paying great attention to the clarity of our proofs. The only real prerequisite for the course is the appropriate mathematical maturity of the students. Although the material found in calculus is independent of that of abstract algebra, a year of calculus is typically given as a prerequisite. Since many of the examples in algebra comes from matrices, we assume that the reader has some basic knowledge of matrix theory. The book should prepare the student for higher level mathematics courses and computer science courses. We have many problems of varying difficulty appearing after each section. We occasionally leave as an exercise the verification of a certain point in a proof. However, we do not rely on exercises to introduce concepts which will be needed later on in the text.

A distinguishing feature of the book is the Worked-Out Exercises which appear after every section. These Worked-Out Exercises provide not only techniques of problem solving, but also supply additional information to enhance the level of knowledge of the reader. The reader should study the Worked-Out Exercises that are marked with $\diamondsuit$ along with the chapter. Those not marked with $\diamondsuit$ may be skipped during the first reading. Sprinkled throughout the book are comments dealing with the historical development of abstract algebra.

We welcome any comments concerning the text. The comments may be forwarded to the following e-mail addresses: malik@creighton.edu or mordes@creighton.edu

D.S. Malik
J. N. Mordeson
M.K. Sen

Contents

List of Symbols

$\in$	belongs to
$\notin$	does not belong to
$\subseteq$	subset
$\subset$	proper subset
$\supseteq$	contains
$\supset$	properly contains
Δ	symmetric difference
$A \backslash B$	set difference
(a, b)	ordered pair
A'	complement of a set A
$\mathbf{N}$	set of positive integers
$\mathbf{Z}$	set of integers
$\mathbf{Z}^{\#}$	set of nonnegative integers
$\mathbf{Q}$	set of rational numbers
$\mathbf{Q}^{+}$	set of positive rational numbers
$\mathbf{Q}^{*}$	set of nonzero rational numbers
$\mathbf{R}$	set of real numbers
$\mathbf{R}^{+}$	set of positive real numbers
$\mathbf{R}^{*}$	set of nonzero real numbers
$\mathbf{C}$	set of complex numbers
$\mathbf{C}^{*}$	set of nonzero complex numbers
$\mathcal{P}(S)$	power set of the set S
$\cup$	union of sets
$\cap$	intersection of sets
$\binom{n}{i}$	number of combinations of n objects taken i at a time
$n!$	n factorial
$a \mid b$	a divides b
$a \nmid b$	a does not divide b
$\gcd(a, b)$	greatest common divisor of a and b
$\operatorname{lcm}(a, b)$	least common multiple of a and b
$\sum_{i=1}^{n} a_i$	$a_1 + a_2 + \cdots + a_n$
$\sum_{a \in S} a$	sum of all elements of S
$\equiv_n$	congruence modulo n
$f : A \to B$	f is a function from a set A into a set B
$f(x)$	image of x under f
$\mathcal{D}(f)$	domain of f
$\mathcal{I}(f)$	image of f
$g \circ f$	composition of mappings g and f
f^{-1}	inverse of a mapping f
I_n	$I_n = \{1, 2, \ldots, n\}$
$f(A)$	$f(A) = \{f(a) \mid a \in A\}$, A is a set contained in the domain of the function f
$f^{-1}(B)$	$f^{-1}(B) = \{x \in X \mid f(x) \in B\}$, where $f : X \to Y$ and $B \subseteq Y$
$\circ$	composition
Π	product

$M_n(R)$	set of all $n \times n$ matrices over R
$\|X\|$	number of elements in a set X
$\|G\|$	order of the group G
$\circ(a)$	order of an element a
$\mathbf{Z}_n$	set of integers modulo n
$Z(G)$	center of the group G
G/H	quotient group
aH, Ha	left, right coset of a in H
aHa^{-1}	$aHa^{-1} = \{aha^{-1} \mid h \in H\}$
$[G : H]$	index of the subgroup H in G
K_4	Klein 4-group
S_n	symmetric group on n symbols
A_n	alternating group on n symbols
D_n	dihedral group of degree n
$\langle S \rangle$	the subgroup generated by S
$\langle a \rangle$	the subgroup generated by a
$\oplus$	direct sum
$N(H)$	Normalizer of H
$C(a)$	centralizer of a
Ker f	kernel of f
$\simeq$	isomorphism
$\mathrm{Aut}(G)$	set of all automorphisms of the group G
$\mathrm{Inn}(G)$	set of all inner automorphisms of the group G
G_a	stabilizer of a or isotropy group of a
$Cl(a)$	conjugacy class of a
G'	commutator subgroup of the group G
$G[n]$	set of all $x \in G$ with $nx = 0$, G is a group
nG	$nG = \{nx \mid x \in G\}$
$C(R)$	center of the ring R
$Q_{\mathbf{R}}$	real quaternions
$\mathbf{Z}[\sqrt{n}]$	$\mathbf{Z}[\sqrt{n}] = \{a + b\sqrt{n} \mid a, b \in \mathbf{Z}\}$, n is a fixed positive integer
$\mathbf{Z}[i]$	$\mathbf{Z}[i] = \{a + bi \mid a, b \in \mathbf{Z}\}$
$\mathbf{Z}[i\sqrt{n}]$	$\mathbf{Z}[i\sqrt{n}] = \{a + bi\sqrt{n} \mid a, b \in \mathbf{Z}\}$, n is a fixed positive integer
$\mathbf{Q}[\sqrt{n}]$	$\mathbf{Q}[\sqrt{n}] = \{a + b\sqrt{n} \mid a, b \in \mathbf{Q}\}$, n is a fixed positive integer
$\mathbf{Q}[i]$	$\mathbf{Q}[i] = \{a + bi \mid a, b \in \mathbf{Q}\}$
$\mathbf{Q}[i\sqrt{n}]$	$\mathbf{Q}[i\sqrt{n}] = \{a + bi\sqrt{n} \mid a, b \in \mathbf{Q}\}$, n is a fixed positive integer
$\langle a \rangle_l$	the left ideal generated by a
$\langle a \rangle_r$	the right ideal generated by a
$\langle a \rangle$	the ideal generated by a
R/I	quotient ring
$Q(R)$	quotient field of the ring R
$R[x]$	polynomial ring in x
$\deg f(x)$	degree of the polynomial $f(x)$
$R[x_1, x_2, \ldots, x_n]$	polynomial ring in n indeterminates
$\sqrt{I}$	radical of an ideal I
rad$R$	Jacobson radical of a ring R
F/K	field extension
$K(C)$	smallest subfield containing the subfield K and the subset C of a field
$[F : K]$	degree of the field F over the field K
$\mathrm{GF}(n)$	Galois field of n elements
$G(F/K)$	Galois group of the field F over the field K
F_G	fixed field of the group G
$\Phi_n(x)$	nth cyclotomic polynomial
P_F	plane of the field F

B^n	set of all binary n-tuples
x^α	$x_1^{\alpha_1} \cdots x_n^{\alpha_n}$
K^n	affine space over the field K
$I(V)$	ideal of the variety V
$\succ$	total ordering
$\succ_l$	lexicographic order
$\succ_{grl}$	graded lexicographic order
$\succ_{grel}$	graded reverse lexicographic order
multideg$f$	multidegree of the polynomial f
LC(f)	leading coefficient of the polynomial f
LM(f)	leading monomial of the polynomial f
LT(f)	leading term of the polynomial f
■	end of proof

Chapter 1

Sets, Relations, and Integers

The purpose of this introductory chapter is mainly to review briefly some familiar properties of sets, functions, and number theory. Although most of these properties are familiar to the reader, there are certain concepts and results which are basic to the understanding of the body of the text.

This chapter is also used to set down the conventions and notations to be used throughout the book. Sets will always be denoted by capital letters. For example, we use the notation $\mathbb{N}$ for the set of positive integers, $\mathbb{Z}$ for the set of integers, $\mathbb{Z}^{\#}$ for the set of nonnegative integers, $\mathbb{E}$ for the set of even integers, $\mathbb{Q}$ for the set of rational numbers, $\mathbb{Q}^{+}$ for the set of positive rational numbers, $\mathbb{Q}^{*}$ for the set of nonzero rational numbers, $\mathbb{R}$ for the set of real numbers, $\mathbb{R}^{+}$ for the set of positive real numbers, $\mathbb{R}^{*}$ for the set of nonzero real numbers, $\mathbb{C}$ for the set of complex numbers, and $\mathbb{C}^{*}$ for the set of nonzero complex numbers.

1.1 Sets

We will not attempt to give an axiomatic treatment of set theory. Rather we use an intuitive approach to the subject. Consequently, we think of a **set** as some given collection of objects. A set S with only a finite number of elements is called a **finite** set; otherwise S is called an **infinite** set. We let $|S|$ denote the number of elements of S. We quite often denote a finite set by a listing of its elements within braces. For example, $\{1, 2, 3\}$ is the set consisting of the objects $1, 2, 3$. This technique is sometimes used for infinite sets. For instance, the set of positive integers $\mathbb{N}$ may be denoted by $\{1, 2, 3, \ldots\}$.

Given a set S, we use the notation $x \in S$ and $x \notin S$ to mean x is a member of S and x is not a member of S, respectively. For the set $S = \{1, 2, 3\}$, we have $1 \in S$ and $4 \notin S$.

A set A is said to be a **subset** of a set S if every element of A is an element of S. In this case, we write $A \subseteq S$ and say that A is contained in S. If $A \subseteq S$, but $A \neq S$, then we write $A \subset S$ and say that A is properly contained in S or that A is a **proper subset** of S. As an example, we have $\{1, 2, 3\} \subseteq \{1, 2, 3\}$ and $\{1, 2\} \subset \{1, 2, 3\}$.

Let A and B be sets. If every member of A is a member of B and every member of B is a member of A, then we say that A and B are the **same** or **equal**. In this case, we write $A = B$. It is immediate that $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$. Thus, we have the following theorem.

Theorem 1.1.1 *Let A and B be sets. Then $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$.* ■

The **null set** or **empty set** is the set with no elements. We usually denote the empty set by $\emptyset$. For any set A, we have $\emptyset \subseteq A$. The later inclusion follows vacuously. That is, every element of $\emptyset$ is an element of A since $\emptyset$ has no elements.

We also describe sets in the following manner. Given a set S, the notation

$$A = \{x \mid x \in S, P(x)\}$$

or

$$A = \{x \in S \mid P(x)\}$$

means that A is the set of all elements x of S such that x satisfies the property P. For example, $\mathbb{N} = \{x \mid x \in \mathbb{Z}, x > 0\}$.

We can combine sets in several ways.

Definition 1.1.2 *The **union** of two sets A and B, written $A \cup B$, is defined to be the set*

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

In the above definition, we mean x is a member of A or x is a member of B or x is a member of both A and B.

Definition 1.1.3 *The **intersection** of two sets A and B, written $A \cap B$, is defined to be the set*

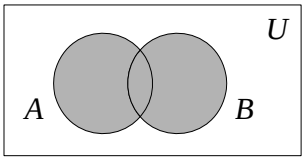
$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

Here x is an element of $A \cap B$ if and only if x is a member of A and at the same time x is a member of B.

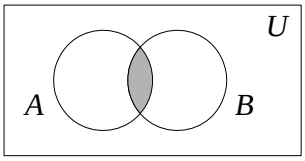
Let A and B be sets. By the definition of the union of sets, every element of A is an element of $A \cup B$. That is, $A \subseteq A \cup B$. Similarly, every element of B is also an element of $A \cup B$ and so $B \subseteq A \cup B$. Also, by the definition of the intersection of sets, every element of $A \cap B$ is an element of A and also an element of B. Hence, $A \cap B \subseteq A$ and $A \cap B \subseteq B$. We record these results in the following theorem.

Theorem 1.1.4 *Let A and B be sets. Then the following statements hold:*
(i) $A \subseteq A \cup B$ and $B \subseteq A \cup B$.
(ii) $A \cap B \subseteq A$ and $A \cap B \subseteq B$. ■

The union and intersection of two sets A and B is described pictorially in the following diagrams. The shaded area represents the set in question.



$A \cup B$



$A \cap B$

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

Example 1.1.5 *Let A be the set $\{1, 2, 3, 4\}$ and B be the set $\{3, 4, 5, 6\}$. Then*

$$A \cup B = \{1, 2, 3, 4, 5, 6\}$$

and $A \cap B = \{3, 4\}$. If C is the set $\{5, 6\}$, then

$$A \cup C = \{1, 2, 3, 4, 5, 6\}$$

while $A \cap C = \emptyset$.

Now that the union and intersection have been defined for two sets, these operations can be similarly defined for any finite number of sets. That is, suppose that $A_1, A_2, \ldots, A_n$ are n sets. The union of $A_1, A_2, \ldots, A_n$, denoted by $\cup_{i=1}^{n} A_i$ or $A_1 \cup A_2 \cup \cdots \cup A_n$, is the set of all elements x such that x is an element of some A_i, where $1 \leq i \leq n$. The intersection of $A_1, A_2, \ldots, A_n$, denoted by $\cap_{i=1}^{n} A_i$ or $A_1 \cap A_2 \cap \cdots \cap A_n$, is the set of all elements x such that $x \in A_i$ for all i, $1 \leq i \leq n$.

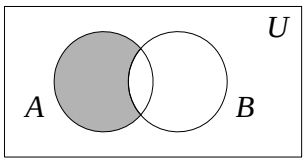
We say that a set I is an **index set** for a collection of sets $\mathcal{A}$ if for any $\alpha \in I$, there exists a set $A_\alpha \in \mathcal{A}$ and $\mathcal{A} = \{A_\alpha \mid \alpha \in I\}$. I can be any nonempty set, finite or infinite.

The **union** of the sets A_α, $\alpha \in I$, is defined to be the set $\{x \mid x \in A_\alpha \text{ for at least one } \alpha \in I\}$ and is denoted by $\cup_{\alpha \in I} A_\alpha$. The **intersection** of the sets A_α, $\alpha \in I$, is defined to be the set $\{x \mid x \in A_\alpha \text{ for all } \alpha \in I\}$ and is denoted by $\cap_{\alpha \in I} A_\alpha$.

Definition 1.1.6 *Given two sets A and B, the* ***relative complement*** *of B in A, denoted by the set difference $A\backslash B$, is the set*

$$A\backslash B = \{x \mid x \in A, \text{ but } x \notin B\}.$$

The following diagram describes the set difference of two sets.



$A\backslash B$

Example 1.1.7 *Let $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$. Then $A\backslash B = \{1, 2\}$.*

We now define a concept which is a building block for all of mathematics, namely, the concept of an ordered pair.

Definition 1.1.8 *Let A and B be nonempty sets and $x \in A$, $y \in B$.*

(i) The ***ordered pair*** *(x, y) is defined to be the set $\{\{x\}, \{x, y\}\}$.*

(ii) The ***Cartesian cross product (Cartesian product)*** *of A and B, written $A \times B$, is defined to be the set*

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

Let $(x, y), (z, w) \in A \times B$. We claim that $(x, y) = (z, w)$ if and only if $x = z$ and $y = w$. First suppose that $x = z$ and $y = w$. Then $\{\{x\}, \{x, y\}\} = \{\{z\}, \{z, w\}\}$ and so $(x, y) = (z, w)$. Now suppose that $(x, y) = (z, w)$. Then

$$\{\{x\}, \{x, y\}\} = \{\{z\}, \{z, w\}\}.$$

Since $\{x\} \in \{\{x\}, \{x, y\}\}$, it follows that $\{x\} \in \{\{z\}, \{z, w\}\}$. This implies that $\{x\} = \{z\}$ or $\{x\} = \{z, w\}$. If $\{x\} = \{z\}$, then we must have $\{x, y\} = \{z, w\}$. From this, it follows that $x = z$ and $y = w$. If $\{x\} = \{z, w\}$, then we must have $\{x, y\} = \{z\}$. This implies that $x = z = w$ and $x = y = z$. Thus, in this case, $x = y = z = w$. This establishes our claim.

It now follows that if A has m elements and B has n elements, then $A \times B$ has mn elements.

Example 1.1.9 *Let $A = \{1, 2, 3\}$ and $B = \{3, 4\}$. Then*

$$A \times B = \{(1, 3), (1, 4), (2, 3), (2, 4), (3, 3), (3, 4)\}.$$

For the set $\mathbb{R}$ of real numbers, the Cartesian product $\mathbb{R} \times \mathbb{R}$ is merely the Euclidean plane.

Definition 1.1.10 *For any set X, the* ***power set*** *of X, written $\mathcal{P}(X)$, is defined to be the set $\{A \mid A$ is a subset of $X\}$.*

Example 1.1.11 *Let $X = \{1, 2, 3\}$. Then*

$$\mathcal{P}(X) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

Here $\mathcal{P}(X)$ has 2^3 elements.

Remark 1.1.12 *Let P and Q be statements. Throughout the text we will encounter questions in which we will be asked to show that P if and only if Q; that is, show that statement P is true if and only if statement Q is true. In situations like this, we first assume that statement P is true and show that statement Q is true. Then we assume that statement Q is true and show that statement P is true. The statement P if and only if Q is also equivalent to the statement: if P, then Q, and if Q, then P. For example, see Worked-Out Exercise 1, below.*

Worked-Out Exercises

Exercise 1 Prove for sets A and B that $A \subseteq B$ if and only if $A \cup B = B$.

Solution First suppose $A \subseteq B$. We now show that $A \cup B = B$. Let x be any element of $A \cup B$. Then either $x \in A$ or $x \in B$. This implies that $x \in B$ since $A \subseteq B$. Thus, we find that every element of $A \cup B$ is an element of B and so $A \cup B \subseteq B$. Also, $B \subseteq A \cup B$ by Theorem 1.1.4(i). Hence, $A \cup B = B$.

Conversely, suppose $A \cup B = B$. Now by Theorem 1.1.4(i), $A \subseteq A \cup B$. Since $A \cup B = B$, it now follows that $A \subseteq B$.

Exercise 2 For a subset A of a set S, let A' denote the subset $S \backslash A$. A' is called the **complement** of A in S. Let A and B be subsets of S. Prove that $(A \cap B)' = A' \cup B'$, **DeMorgan's law**.

Solution First we show that $(A \cap B)' \subseteq A' \cup B'$. Then we show that $A' \cup B' \subseteq (A \cap B)'$. The result then follows by Theorem 1.1.1.

Let x be any element of $(A \cap B)'$. Now $(A \cap B)' = S \backslash (A \cap B)$ and so $x \in S$ and $x \notin A \cap B$. Also, $x \notin A \cap B$ implies that either $x \notin A$ or $x \notin B$. If $x \in S$ and $x \notin A$, then $x \in A'$, and if $x \in S$ and $x \notin B$, then $x \in B'$. Thus, either $x \in A'$ or $x \in B'$, i.e., $x \in A' \cup B'$. Hence, $(A \cap B)' \subseteq A' \cup B'$.

Let us now show that $A' \cup B' \subseteq (A \cap B)'$. Suppose x is any element of $A' \cup B'$. Then either $x \in A'$ or $x \in B'$. Suppose $x \in A'$, then $x \in S$ and $x \notin A$. Since $A \cap B \subseteq A$ and $x \notin A$, we must have $x \notin A \cap B$. This implies that $x \in (A \cap B)'$. Similarly, we can show that if $x \in B'$, then $x \notin A \cap B$, i.e., $x \in (A \cap B)'$. Hence, $A' \cup B' \subseteq (A \cap B)'$. Consequently, $(A \cap B)' = A' \cup B'$.

Exercise 3 Let A, B, and C be sets. Prove that

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

Solution As in the previous exercise, we first show that $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$ and then $(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C)$. The result then follows by Theorem 1.1.1.

Let x be any element of $A \cap (B \cup C)$. Then $x \in A$ and $x \in B \cup C$. Thus, $x \in A$ and $x \in B$ or $x \in C$. If $x \in A$ and $x \in B$, then $x \in A \cap B$, and if $x \in A$ and $x \in C$, then $x \in A \cap C$. Therefore, $x \in A \cap B$ or $x \in A \cap C$. Hence, $x \in (A \cap B) \cup (A \cap C)$. This shows that $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$.

Let us now show that $(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C)$. Suppose x is any element of $(A \cap B) \cup (A \cap C)$. Then $x \in A \cap B$ or $x \in A \cap C$. Suppose $x \in A \cap B$, then $x \in A$ and $x \in B$. Since $B \subseteq B \cup C$, we have $x \in B \cup C$. Thus, $x \in A$ and $x \in B \cup C$ and so $x \in A \cap (B \cup C)$. Similarly, if $x \in A$ and $x \in C$, then $x \in A \cap (B \cup C)$. Hence, $(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C)$. Consequently, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Exercises

1. Let $A = \{x, y, z\}$ and $B = \{y, w\}$. Determine each of the following sets: $A \cup B$, $A \cap B$, $A \backslash B$, $B \backslash A$, $A \times B$, and $\mathcal{P}(A)$.

2. Prove for sets A and B that $A \subseteq B$ if and only if $A \cap B = A$.

3. Prove for sets A, B, and C that

   (i) $A \cup B = B \cup A$ and $A \cap B = B \cap A$,

   (ii) $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$,

   (iii) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$,

   (iv) $A \cup (A \cap B) = A$,

   (v) $A \cap (A \cup B) = A$.

4. If a set S has 12 elements, how many elements does $\mathcal{P}(S)$ have? How many of these are properly contained in S?

5. For subsets A and B of a set S, prove **DeMorgan's law**:

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

6. The **symmetric difference** of two sets A and B is the set

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

(i) If $A = \{a, b, c\}$ and $B = \{b, c, d, e\}$, find $A \triangle B$.

(ii) Show that $A \triangle B = (A \backslash B) \cup (B \backslash A)$.

7. Let A and B be finite subsets of a set S. Show that

(i) if $A \cap B = \emptyset$, then $|A \cup B| = |A| + |B|$,

(ii) $|A \backslash B| = |A| - |A \cap B|$,

(iii) $|A \cup B| = |A| + |B| - |A \cap B|$.

8. In each of the following exercises, write the proof if the statement is true; otherwise give a counterexample. The sets A, B, and C are subsets of a set U.

(i) $A \cap (B \backslash C) = (A \cap B) \backslash (A \cap C)$.

(ii) $A \backslash (B \cup C) = (A \backslash B) \cup C$.

(iii) $(A \backslash B)' = (B \backslash A)'$.

(iv) $A \times (B \cup C) = (A \times B) \cup (A \times C)$.

(v) $A \triangle C = B \triangle C$ implies $A = B$.

1.2 Integers

Throughout abstract algebra, the set of **integers** provides a source of examples. In fact, many algebraic abstractions come from the integers. An axiomatic development of the integers is not given in this text. Instead, certain basic properties of integers are taken for granted. For example, if n and m are integers with $n < m$, then there exists a positive integer $t \in \mathbb{Z}$ such that $m = n + t$. In this section, we review and prove some important properties of the integers.

The proofs of many results of algebra depend on the following basic principle of the integers.

Principle of Well-Ordering: Every nonempty subset of $\mathbb{Z}^{\#}$ has a smallest (least) element, i.e., if $\emptyset \neq S \subseteq \mathbb{Z}^{\#}$, then there exists $x \in S$ such that $x \leq y$ for all $y \in S$.

Let S be a subset of $\mathbb{Z}^{\#}$. Suppose that S has the following properties:
(i) $n_0 \in S$, i.e., there exists an element $n_0 \in S$.
(ii) For all $n \geq n_0$, $n \in \mathbb{Z}^{\#}$, if $n \in S$, then $n + 1 \in S$.
We show that the set of all integers greater than or equal to n_0 is a subset of S, i.e.,

$$\{n \in \mathbb{Z}^{\#} \mid n \geq n_0\} \subseteq S.$$

Let T denote the set $\{n \in \mathbb{Z}^{\#} \mid n \geq n_0\}$. We wish to show that $T \subseteq S$. On the contrary, suppose $T \not\subseteq S$. Then there exists $a \in T$ such that $a \notin S$. Let T_1 be the set of all elements of T that are not in S, i.e., $T_1 = T \backslash S$. Since $a \in T$ and $a \notin S$, we have $a \in T_1$. Thus, T_1 is a nonempty subset of $\mathbb{Z}^{\#}$. Hence, by the principle of well-ordering, T_1 has a smallest element m, say. Then $m \in T$ and $m \notin S$. Since $m \in T$, $m \geq n_0$. If $m = n_0$, then $m \in S$, a contradiction. Thus, $m > n_0$. This implies that $m - 1 \geq n_0$ and so $m - 1 \in T$. Now $m - 1 \notin T_1$ since m is the smallest element of T_1. Since $m - 1 \in T$ and $m - 1 \notin T_1$, we must have $m - 1 \in S$. But then by (ii), $m = (m - 1) + 1 \in S$, which is a contradiction. Hence, $T \subseteq S$.

Thus, from the principle of well-ordering, we deduce another important property of integers. This property is known as the principle of mathematical induction. We thus have the following theorem.

Theorem 1.2.1 (Principle of Mathematical Induction) *Let $S \subseteq \mathbb{Z}^{\#}$. Let $n_0 \in S$. Suppose S satisfies either of the following conditions.*

(i) For all $n \geq n_0$, $n \in \mathbb{Z}^{\#}$, if $n \in S$, then $n+1 \in S$.

(ii) For all $n_0 \leq m < n$, $n \in \mathbb{Z}^{\#}$, if $m \in S$, then $n \in S$.

Then

$$\{n \in \mathbb{Z}^{\#} \mid n \geq n_0\} \subseteq S. \blacksquare$$

We proved, above, Theorem 1.2.1, when S satisfies (i). We leave it for the reader to prove Theorem 1.2.1 if S satisfies (ii).

We have seen the following mathematical statement in a college algebra or in a calculus course.

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}, \quad n \geq 1.$$

We now show how this statement can be proved using the principle of mathematical induction. Let $S(n)$ denote the above mathematical statement, i.e.,

$$S(n): \quad 1 + 2 + \cdots + n = \frac{n(n+1)}{2}, \quad n \geq 1.$$

This statement will be true if the left-hand side of the statement is equal to the right-hand side. Let

$$S = \{n \in \mathbb{Z}^{\#} \mid S(n) \text{ is true}\}.$$

That is, S is the set of all nonnegative integers n for which the statement $S(n)$ is true. We will show that S is the set of all positive integers. Now

$$1 = \frac{1 \cdot (1+1)}{2},$$

i.e., $S(1)$ is true. Hence, $1 \in S$. Let n be an integer such that $n \geq 1$ and suppose $S(n)$ is true, i.e., $n \in S$. We now show that $S(n+1)$ is true. Now

$$S(n+1): \quad 1 + 2 + \cdots + n + (n+1) = \frac{(n+1)(n+2)}{2}.$$

Consider the left-hand side.

$$\begin{aligned} 1 + 2 + \cdots + n + (n+1) &= \tfrac{n(n+1)}{2} + (n+1) \text{ (since } S(n) \text{ is true)} \\ &= \tfrac{(n+1)(n+2)}{2}. \end{aligned}$$

Hence, the left-hand side is equal to the right-hand side and so $S(n+1)$ is true. Thus, $n+1 \in S$. Hence, by the principle of mathematical induction, $S = \{n \in \mathbb{Z}^{\#} \mid n \geq 1\}$. This proves our claim, which in turn shows that

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

is true for all positive integers n.

Sometimes we use the word induction for the principle of mathematical induction.

A proof by the principle of mathematical induction consists of three steps.

Step 1: Show that $n_0 \in S$, i.e., the statement $S(n_0)$ is true for some $n_0 \in \mathbb{Z}^{\#}$.

Step 2: Write the induction hypothesis: n is an integer such that $n \geq n_0$ and $n \in S$, i.e., $S(n)$ is true for some integer n such that $n \geq n_0$ (or k is an integer such that $n_0 \leq k \leq n$ and $S(k)$ is true).

Step 3: Show that $n+1 \in S$, i.e., $S(n+1)$ is true.

Example 1.2.2 *In this example, we show that* $2n+1 \leq 2^n$ *for all* $n \geq 3$.

Let $S(n)$ *be the statement:*

$$S(n): \quad 2n+1 \leq 2^n, \quad n \geq 3.$$

Since we want to show that $S(n)$ *is true for all* $n \geq 3$, *as the first step of our induction, we must verify that* $S(3)$ *is true. Let* $n = 3$. *Now* $2n+1 = 2 \cdot 3 + 1 = 7$ *and* $2^n = 2^3 = 8$. *Thus, for* $n = 3$, $2n+1 \leq 2^n$. *This shows that* $S(3)$ *is true. Suppose that* $2n+1 \leq 2^n$ *for some* $n \geq 3$, *i.e.,* $S(n)$ *is true for some* $n \geq 3$. *Consider* $S(n+1)$,

$$S(n+1): \quad 2(n+1)+1 \leq 2^{n+1}.$$

Let us evaluate the left-hand side of $S(n+1)$. *We have*

$$\begin{aligned} 2(n+1)+1 &= 2n+2+1 \\ &= (2n+1)+2 \\ &\leq 2^n + 2 \qquad \textit{since } S(n) \textit{ is true} \\ &\leq 2^n + 2^n \qquad (\textit{since } n \geq 3,\ 2 \leq 2^n) \\ &= 2^{n+1}. \end{aligned}$$

Thus, $S(n+1)$ *is true. Hence, by the principle of mathematical induction,* $2n+1 \leq 2^n$ *for all* $n \geq 3$.

The principle of mathematical induction is a very useful tool in mathematics. We will make use of this result throughout the text.

We now prove the following important properties of integers with the help of the principle of well-ordering.

Theorem 1.2.3 (Division Algorithm) *Let* $x, y \in \mathbb{Z}$ *with* $y \neq 0$. *Then there exist unique integers* q *and* r *such that* $x = qy + r$, $0 \leq r < |y|$.

Proof. Let us first assume $y > 0$. Then $y \geq 1$. Consider the set

$$S = \{x - uy \mid u \in \mathbb{Z}, x - uy \geq 0\}.$$

Since $y \geq 1$, we have $x - (-|x|)y = x + |x|\, y \geq 0$ so that $x - (-|x|)y \in S$. Thus, S is a nonempty set of nonnegative integers. Hence, by the principle of well-ordering, S must have a smallest element, say, r. Since $r \in S$, we have $r \geq 0$ and $r = x - qy$ for some $q \in \mathbb{Z}$. Then $x = qy + r$. We must show that $r < |y|$. Suppose on the contrary that $r \geq |y| = y$. Then

$$x - (q+1)y = (x - qy) - y = r - y \geq 0$$

so that $r - y \in S$, a contradiction since r is the smallest nonnegative integer in S and $r - y < r$. Hence, it must be the case that $r < |y|$. This proves the theorem in case $y > 0$.

Suppose now that $y < 0$. Then $|y| > 0$. Thus, there exist integers q', r such that $x = q'|y| + r$, $0 \leq r < |y|$ by the above argument. Since $y < 0$, $|y| = -y$. Hence, $x = -q'y + r$. Let $q = -q'$. Then $x = qy + r$, $0 \leq r < |y|$, the desired conclusion.

The uniqueness of q and r remains to be shown. Suppose there are integers q', r' such that

$$x = qy + r = q'y + r',$$

$0 \leq r' < |y|$, $0 \leq r < |y|$. Then

$$r' - r = (q - q')y.$$

Thus,

$$|r' - r| = |q - q'|\, |y|.$$

Now $-|y| < -r \leq 0$ and $0 \leq r' < |y|$. Therefore, if we add these inequalities, we obtain

$$-|y| < r' - r < |y|,$$

or $|r' - r| < |y|$. Hence, we have

$$0 \leq |q - q'| < 1.$$

Since $q - q'$ is an integer, we must have $0 = |q - q'|$. It now also follows that $|r - r'| = 0$. Thus, $q - q' = 0$ and $r - r' = 0$ or $q = q'$ and $r = r'$. Consequently, q and r are unique. ■

In Theorem 1.2.3, the integer q is called the **quotient** of x and y on dividing x by y and the integer r is called the **remainder** of x and y on dividing x by y.

The following corollary is a special case of Theorem 1.2.3.

Corollary 1.2.4 *For any two integers x and y with $y > 0$, there exist unique integers q and r such that $x = qy + r$, where $0 \leq r < y$.*

Proof. By Theorem 1.2.3, there exist unique integers q and r such that $x = qy + r$, where $0 \leq r < |y|$. Since $y > 0$, $|y| = y$. Hence, $x = qy + r$, where $0 \leq r < y$. ■

Definition 1.2.5 *Let $x, y \in \mathbb{Z}$ with $x \neq 0$. Then x is said to **divide** y or x is a **divisor** (or **factor**) of y, written $x|y$, provided there exists $q \in \mathbb{Z}$ such that $y = qx$. When x does not divide y, we sometimes write $x \nmid y$.*

Let x, y, z be integers with $x \neq 0$. Suppose $x|y$ and $x|z$. Then for all integers s and t, $x|(sy + tz)$. We ask the reader to prove this fact in Exercise 5(iii) (page 16).

Definition 1.2.6 *Let $x, y \in \mathbb{Z}$. A nonzero integer c is called a **common divisor** of x and y if $c|x$ and $c|y$.*

Definition 1.2.7 *A nonzero integer d is called a **greatest common divisor (gcd)** of the integers x and y if*

(i) $d|x$ and $d|y$,

(ii) for all $c \in \mathbb{Z}$ if $c|x$ and $c|y$, then $c|d$.

Let d and d' be two greatest common divisors of integers x and y. Then $d|d'$ and $d'|d$. Hence, there exist integers u and v such that $d' = du$ and $d = d'v$. Therefore, $d = duv$, which implies that $uv = 1$ since $d \neq 0$. Thus, either $u = v = 1$ or $u = v = -1$. Hence, $d' = \pm d$. It now follows that two different gcd's of x and y differ in their sign. Of the two gcd's of x and y, the positive one is denoted by $\gcd(x, y)$. For example, 2 and -2 are the greatest common divisors of 4 and 6. Hence, $2 = \gcd(4, 6)$.

In the next theorem, we show that the gcd always exists for any two nonzero integers.

Theorem 1.2.8 *Let $x, y \in \mathbb{Z}$ with either $x \neq 0$ or $y \neq 0$. Then x and y have a positive greatest common divisor d. Moreover, there exist elements $s, t \in \mathbb{Z}$ such that $d = sx + ty$.*

Proof. Let

$$S = \{mx + ny \mid m, n \in \mathbb{Z}, mx + ny > 0\}.$$

Suppose $x \neq 0$. Then

$$\begin{aligned} |x| &= \begin{cases} x & \text{if } x > 0 \\ -x & \text{if } x < 0 \end{cases} \\ &= \begin{cases} 1x + 0y & \text{if } x > 0 \\ (-1)x + 0y & \text{if } x < 0. \end{cases} \end{aligned}$$

Hence, $|x| \in S$ and so $S \neq \emptyset$. By the well-ordering principle, S contains a smallest positive integer, say, d. We now show that d is the greatest common divisor of x and y.

Since $d \in S$, there exist $s, t \in \mathbb{Z}$ such that $d = sx + ty$. First we show that $d|x$ and $d|y$. Since $d \neq 0$, by the division algorithm (Theorem 1.2.3), there exist integers q and r such that

$$x = dq + r,$$

where $0 \leq r < |d| = d$. Thus,

$$\begin{aligned} r &= x - dq \\ &= x - (sx + ty)q \qquad \text{(substituting for } d\text{)} \\ &= (1 - qs)x + (-qt)y. \end{aligned}$$

Suppose $r > 0$. Then $r \in S$, which is a contradiction since d is the smallest element of S and $r < d$. Thus, $r = 0$. This implies that $x = dq$ and so $d|x$. Similarly, $d|y$. Hence, d satisfies (i) of Definition 1.2.7. Suppose $c|x$ and $c|y$ for some integer c. Then $c|(sx + ty)$ by Exercise 5(iii) (page 16), i.e., $c|d$. Thus, d satisfies (ii) of Definition 1.2.7. Consequently, $d = \gcd(x, y)$. ■

Let x and y be nonzero integers. By Theorem 1.2.8, $\gcd(x, y)$ exists and if $d = \gcd(x, y)$, then there exist integers s and t such that $d = sx + ty$. The integers s and t in the representation $d = sx + ty$ are not unique. For example, let $x = 45$ and $y = 126$. Then $\gcd(x, y) = 9$, and $9 = 3 \cdot 45 + (-1) \cdot 126 = 129 \cdot 45 + (-46) \cdot 126$.

The proof of Theorem 1.2.8 does not indicate how to find $\gcd(x, y)$ or the integers s, t. In the following, we indicate how these integers can be found.

Let $x, y \in \mathbb{Z}$ with $y \neq 0$. By the division algorithm, there exist $q_1, r_1 \in \mathbb{Z}$ such that

$$x = q_1 y + r_1, \qquad 0 \leq r_1 < |y|.$$

If $r_1 \neq 0$, then by the division algorithm, there exist q_2, $r_2 \in \mathbb{Z}$ such that

$$y = q_2 r_1 + r_2, \qquad 0 \leq r_2 < r_1.$$

If $r_2 \neq 0$, then again by the division algorithm, there exist q_3, $r_3 \in \mathbb{Z}$ such that

$$r_1 = q_3 r_2 + r_3, \qquad 0 \leq r_3 < r_2.$$

Since $r_1 > r_2 > r_3 \geq 0$, we must in a finite number of steps find integers q_n, q_{n+1}, and $r_n > 0$ such that

$$\begin{aligned} r_{n-2} &= q_n r_{n-1} + r_n, \quad 0 < r_n < r_{n-1} \\ r_{n-1} &= q_{n+1} r_n + 0. \end{aligned}$$

We assert that r_n (the last nonzero remainder) is the greatest common divisor of x and y. Now $r_n | r_{n-1}$. Since $r_n | r_n$, $r_n | r_{n-1}$, and $r_{n-2} = q_n r_{n-1} + r_n$, we have $r_n | r_{n-2}$ by Exercise 5(iii) (page 16). Working our way back in this fashion, we have $r_n | r_1$ and $r_n | r_2$. Thus, $r_n | y$ since $y = q_2 r_1 + r_2$. Since $r_n | y$, $r_n | r_1$, and $x = q_1 y + r_1$, we have $r_n | x$. Hence, r_n is a common divisor of x and y. Now if c is any common divisor of x and y, then we see that $c | r_1$. Since $c | y$ and $c | r_1$, $c | r_2$. Continuing, we finally obtain $c | r_n$. Thus, $r_n = \gcd(x, y)$.

We now find s, $t \in \mathbb{Z}$ such that $\gcd(x, y) = sx + ty$ as follows:

$$\begin{aligned} r_n &= r_{n-2} + r_{n-1}(-q_n) \\ &= r_{n-2} + [r_{n-3} + r_{n-2}(-q_{n-1})](-q_n) \\ &= r_{n-3}(-q_n) + r_{n-2}(1 + q_{n-1} q_n) \qquad \text{(simplifying)}. \end{aligned}$$

We now substitute $r_{n-4} + r_{n-3}(-q_{n-2})$ for r_{n-2}. We repeat this "back" substitution process until we reach $r_n = sx + ty$ for some integers s and t.

We illustrate the above procedure for finding the gcd and integers s and t with the help of the following example.

Example 1.2.9 *Consider the integers* 45 *and* 126. *Now*

$$\begin{aligned} 126 &= 2 \cdot 45 + 36 \\ 45 &= 1 \cdot 36 + 9 \\ 36 &= 4 \cdot 9 + 0 \end{aligned}$$

Thus, $9 = \gcd(45, 126)$. *Also,*

$$\begin{aligned} 9 &= 45 - 1 \cdot 36 \\ &= 45 - 1 \cdot [126 - 2 \cdot 45] \\ &= 3 \cdot 45 + (-1) \cdot 126. \end{aligned}$$

Here $s = 3$ *and* $t = -1$.

We now define prime integers and study their basic properties.

Definition 1.2.10 *(i) An integer* $p > 1$ *is called* ***prime*** *if the only divisors of* p *are* ± 1 *and* $\pm p$.
(ii) Two integers x *and* y *are called* ***relatively prime*** *if* $\gcd(x, y) = 1$.

The following theorem gives a necessary and sufficient condition for two nonzero integers to be relatively prime.

Theorem 1.2.11 *Let* x *and* y *be nonzero integers. Then* x *and* y *are relatively prime if and only if there exist* $s, t \in \mathbb{Z}$ *such that* $1 = sx + ty$.

Proof. Let x and y be relatively prime. Then $\gcd(x, y) = 1$. By Theorem 1.2.8, there exist integers s and t such that $1 = sx + ty$.

Conversely, suppose $1 = sx + ty$ for some pair of integers s, t. Let $d = \gcd(x, y)$. Then $d|x$ and $d|y$ and so $d|(sx+ty)$ (by Exercise 5(iii) (page 16)) or $d|1$. Since d is a positive integer and $d|1$, $d = 1$. Thus, $\gcd(x, y) = 1$ and so x and y are relatively prime. ■

Theorem 1.2.12 *Let* x, y, $z \in \mathbb{Z}$ *with* $x \neq 0$. *If* $x|yz$ *and* x, y *are relatively prime, then* $x|z$.

Proof. Since x and y are relatively prime, there exist $s, t \in \mathbb{Z}$ such that $1 = sx + ty$ by Theorem 1.2.11. Thus, $z = sxz + tyz$. Now $x|x$ and by hypothesis $x|yz$. Thus, $x|(sxz + tyz)$ by Exercise 5(iii) (page 16) and so $x|z$. ■

Corollary 1.2.13 *Let* $x, y, p \in \mathbb{Z}$ *with* p *a prime. If* $p|xy$, *then either* $p|x$ *or* $p|y$.

Proof. If $p|x$, then we have the desired result. Suppose that p does not divide x. Since the only positive divisors of p are 1 and p, we must have that p and x are relatively prime. Thus, $p|y$ by Theorem 1.2.12. ■

The following corollary is a generalization of Corollary 1.2.13.

Corollary 1.2.14 *Let* $x_1, x_2, \ldots, x_n, p \in \mathbb{Z}$ *with* p *a prime. If*

$$p|x_1 x_2 \cdots x_n,$$

then $p|x_i$ *for some* i, $1 \leq i \leq n$.

Proof. The proof follows by Corollary 1.2.13 and induction. ■

Consider the integer 24. We can write $24 = 2^3 \cdot 3$. That is, 24 can be written as product of prime powers. Similarly, $49500 = 2^2 \cdot 3^2 \cdot 5^3 \cdot 11$. In the next theorem, called the fundamental theorem of arithmetic, we prove that any positive integer can be written as product of prime powers.

Theorem 1.2.15 (Fundamental Theorem of Arithmetic) *Any integer* $n > 1$ *has a unique factorization (up to order)*

$$n = p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s}, \tag{1.1}$$

where $p_1, p_2, \ldots, p_s$ *are distinct primes and* $e_1, e_2, \ldots, e_s$ *are positive integers.*

Proof. First we show that any integer $n > 1$ has a factorization like Eq. (1.1) and then we show the uniqueness of the factorization.

We show the existence of the factorization by induction. If $n = 2$, then clearly n has the above factorization as a product of prime powers. Make the induction hypothesis that any integer k such that $2 \leq k < n$ has a factorization like Eq. (1.1). If n is prime, then n already has the above factorization as a product of prime powers, namely n itself. If n is not prime, then $n = xy$ for integers x, y, with $1 < x < n$ and $1 < y < n$. By the induction hypothesis, there exist primes $q_1, q_2, \ldots, q_k, q_1', q_2', \ldots, q_t'$ and positive integers $e_1, e_2, \ldots, e_k, e_1', e_2', \ldots, e_t'$ such that $q_1, q_2, \ldots, q_k$ are distinct primes, $q_1', q_2', \ldots, q_t'$ are distinct primes and

$$\begin{aligned} x &= q_1^{e_1} q_2^{e_2} \cdots q_k^{e_k} \\ y &= q_1'^{e_1'} q_2'^{e_2'} \cdots q_t'^{e_t'}. \end{aligned}$$

Thus,

$$n = q_1^{e_1} q_2^{e_2} \cdots q_k^{e_k} q_1'^{e_1'} q_2'^{e_2'} \cdots q_t'^{e_t'},$$

i.e., n can be factored as a product of prime powers. If $q_i = q_j'$ for some i and j, then we replace $q_i^{e_i} q_j'^{e_j'}$ by $q_i^{e_i + e_j'}$. It now follows that $n = p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s}$, where $p_1, p_2, \ldots, p_s$ are distinct primes and $e_1, e_2, \ldots, e_s$ are positive integers. Hence, by induction, any integer $n > 1$ has a factorization like (1.1).

We now prove the uniqueness property by induction also. If $n = 2$, then clearly n has a unique factorization as a product of prime powers. Suppose the uniqueness property holds for all integers k such that $2 \leq k < n$. Let

$$n = p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s} = q_1^{c_1} q_2^{c_2} \cdots q_t^{c_t} \tag{1.2}$$

be two factorizations of n into a product of prime powers. Suppose n is prime. Then in Eq. (1.2), we must have $s = t = 1$ and $e_1 = 1 = c_1$ since the only positive divisors of n are 1 and n itself. This implies that $n = p_1 = q_1$ and so the factorization is unique.

Suppose n is not a prime. Now $p_1 | n$ and

$$\frac{n}{p_1} = p_1^{e_1 - 1} p_2^{e_2} \cdots p_s^{e_s}$$

is an integer. If $s = 1$, then $n = p_1^{e_1}$ and since n is not a prime, we have $e_1 > 1$. Hence, $\frac{n}{p_1} = p_1^{e_1 - 1} \geq 2$. If $s > 1$, then $\frac{n}{p_1} = p_1^{e_1 - 1} p_2^{e_2} \cdots p_s^{e_s} \geq 2$. Thus, in either case, $\frac{n}{p_1}$ is an integer ≥ 2. Now $p_1 | n$ implies that $p_1 | q_1^{c_1} q_2^{c_2} \cdots q_t^{c_t}$ and so by Corollary 1.2.14, $p_1 | q_i^{c_i}$ for some i. By reordering the q_i if necessary, we can assume that $i = 1$. Thus, $p_1 | q_1^{c_1}$ and so by Corollary 1.2.14, $p_1 | q_1$. Since p_1 and q_1 are primes, $p_1 = q_1$. Thus,

$$\frac{n}{p_1} = p_1^{e_1 - 1} p_2^{e_2} \cdots p_s^{e_s} = p_1^{c_1 - 1} q_2^{c_2} \cdots q_t^{c_t}. \tag{1.3}$$

Now $e_1 - 1 = 0$ if and only if $c_1 - 1 = 0$. For suppose $e_1 - 1 = 0$ and $c_1 - 1 > 0$. Then $\frac{n}{p_1} = p_2^{e_2} \cdots p_s^{e_s}$ implies that $p_1 \not| \frac{n}{p_1}$ and $\frac{n}{p_1} = p_1^{c_1 - 1} q_2^{c_2} \cdots q_t^{c_t}$ implies that $p_1 | \frac{n}{p_1}$, which is of course impossible. We can get a similar contradiction if we assume $e_1 - 1 > 0$ and $c_1 - 1 = 0$.

Now $\frac{n}{p_1}$ is an integer and $2 \leq \frac{n}{p_1} < n$. Hence, by the induction hypothesis, we obtain from Eq. (1.3) that $s = t$, and $p_1 = q_1, \ldots, p_s = q_s$ (without worrying about the order), and $e_1 - 1 = c_1 - 1$, $e_2 = c_2, \ldots, e_s = c_s$. Hence, by induction, we have the desired uniqueness property. ■

Corollary 1.2.16 *Any integer $n < -1$ has a unique factorization (up to order)*

$$n = (-1) p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s},$$

where $p_1, p_2, \ldots, p_s$ are distinct primes and $e_1, e_2, \ldots, e_s$ are positive integers.

Proof. Since $n < -1$, $-n > 1$. Hence, by Theorem 1.2.15, $-n$ has a unique factorization (up to order)

$$-n = p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s},$$

where $p_1, p_2, \ldots, p_s$ are distinct primes and $e_1, e_2, \ldots, e_s$ are positive integers. Thus,

$$n = (-1)p_1^{e_1}p_2^{e_2}\cdots p_s^{e_s},$$

where $p_1, \ldots, p_s$ are distinct primes and $e_1, \ldots, e_s$ are positive integers. ■

Theorem 1.2.15 says that any positive integer greater than 1 can be written as a product of prime powers. Now we pose the obvious question: How many prime numbers are there? This is answered by the following theorem due to Euclid.

Theorem 1.2.17 (Euclid) *There are an infinite number of primes.*

Proof. Let $p_1, p_2, \ldots, p_n$ be a finite number of distinct primes. Set $x = p_1p_2\cdots p_n + 1$. Since p_i does not divide 1, p_i does not divide x, $i = 1, 2, \ldots, n$. By the fundamental theorem of arithmetic, it follows that there is some prime p such that $p|x$. Thus, p is distinct from $p_1, p_2, \ldots, p_n$ so that we have $n + 1$ distinct primes. That is, for any finite set of primes we can always find one more. Thus, there must be an infinite number of primes. ■

We close this section with the following definition. There are a few places in the text where we will be making use of it.

Definition 1.2.18 *Let n be a positive integer. Let $\phi(n)$ denote the number of positive integers m such that $m \leq n$ and $\gcd(m, n) = 1$, i.e.,*

$$\phi(n) = |\{m \in \mathbb{N} \mid m \leq n \text{ and } \gcd(m, n) = 1\}|.$$

*$\phi(n)$ is called the **Euler ϕ-function.***

Clearly $\phi(2) = 1$, $\phi(3) = 2$, $\phi(4) = 2$. Since 1, 5, 7, 11 are the only positive integers less than 12 and relatively prime to 12, $\phi(12) = 4$.

Let $\{a_1, \ldots, a_n\} \subseteq \mathbb{Z}$. We use the notation $\sum_{i=1}^{n} a_i$ to denote the sum of $a_1, \ldots, a_n$, i.e.,

$$\sum_{i=1}^{n} a_i = a_1 + \cdots + a_n.$$

If S is any finite subset of $\mathbb{Z}$, then $\sum_{a\in S} a$ denotes the sum of all elements of S. For example, if $S = \{2, 4, 7\}$, then $\sum_{a\in S} a = 2 + 4 + 7 = 13$.

Worked-Out Exercises

Exercise 1 By the principle of mathematical induction, prove that

$$3^{2n+1} + (-1)^n 2 \equiv 0 (\text{mod } 5)$$

for all positive integers n. (For integers a and b, $a \equiv b (\text{mod } 5)$ means 5 divides $a - b$.)

Solution Let $S(n)$ be the statement

$$S(n): \qquad 3^{2n+1} + (-1)^n 2 \equiv 0 (\text{mod } 5), \qquad n \geq 1.$$

We wish to show that $S(n)$ is true for all positive integers. We first must verify that $S(1)$ is true as the first step of our induction. Let $n = 1$. Then

$$3^{2n+1} + (-1)^n 2 = 3^{2+1} + (-1)2 = 27 - 2 = 25 \equiv 0 (\text{mod } 5).$$

Thus, $S(1)$ is true. Now suppose that $S(n)$ is true for some positive integer n, i.e., $3^{2n+1} + (-1)^n 2 \equiv 0 (\text{mod } 5)$ for some integer $n \geq 1$. We now show that

$$S(n+1): \qquad 3^{2(n+1)+1} + (-1)^{n+1} 2 \equiv 0 (\text{mod } 5)$$

is true. Now

$$\begin{aligned} 3^{2(n+1)+1} + (-1)^{n+1}2 &= 3^{2n+1} \cdot 3^2 - (-1)^n 2 \\ &= 9(3^{2n+1} + (-1)^n 2) - (-1)^n 18 - (-1)^n 2 \\ &= 9(3^{2n+1} + (-1)^n 2) - (-1)^n 20. \end{aligned}$$

Since $3^{2n+1} + (-1)^n 2 \equiv 0 (\text{mod } 5)$ and $20 \equiv 0 (\text{mod } 5)$, it follows that $3^{2(n+1)+1} + (-1)^{n+1}2 \equiv 0 (\text{mod } 5)$. This shows that $S(n+1)$ is true. Hence, by the principle of mathematical induction, $3^{2n+1} + (-1)^n 2 \equiv 0 (\text{mod } 5)$ for all positive integers n.

Exercise 2 Let a and b be integers such that $\gcd(a, 4) = 2$ and $\gcd(b, 4) = 2$. Prove that $\gcd(a + b, 4) = 4$.

Solution Since $\gcd(a, 4) = 2$, $2|a$, but 4 does not divide a. Therefore, $a = 2x$ for some integer x such that $\gcd(2, x) = 1$. Similarly, $b = 2y$ for some integer y such that $\gcd(2, y) = 1$. Thus, x and y are both odd integers. This implies that $x + y$ is an even integer and so $x + y = 2n$ for some integer n. Now $a + b = 2(x + y) = 4n$. Hence, $\gcd(a + b, 4) = \gcd(4n, 4) = 4$.

Exercise 3 Let a, b, and c be integers such that $\gcd(a, c) = \gcd(b, c) = 1$. Prove that $\gcd(ab, c) = 1$.

Solution If $c = 0$, then $\gcd(a, 0) = \gcd(b, 0) = 1$ implies that $a = \pm 1$ and $b = \pm 1$. Thus, $\gcd(ab, c) = \gcd(\pm 1, 0) = 1$. Suppose now $c \neq 0$. By Theorem 1.2.8, $\gcd(ab, c)$ exists. Let $d = \gcd(ab, c)$. Also, by Theorem 1.2.8, there exist integers x_1, y_1, x_2, y_2 such that $1 = ax_1 + cy_1$, $1 = bx_2 + cy_2$. Thus, $(ax_1)(bx_2) = (1 - cy_1)(1 - cy_2) = 1 - cy_1 - cy_2 + cy_1cy_2$. Hence, $1 = (ab)x_1x_2 + c(y_1 + y_2 - cy_1y_2)$. Thus, any common divisor of ab and c is also a divisor of 1. Hence, $d|1$. Since $d > 0$, $d = 1$.

Exercise 4 Let $a, b \in \mathbb{Z}$ with either $a \neq 0$ or $b \neq 0$. Prove that for any integer c,

$$\gcd(a, b) = \gcd(a, -b) = \gcd(a, b + ac).$$

Solution Suppose $a \neq 0$. Then $\gcd(a, b)$, $\gcd(a, -b)$ and $\gcd(a, b + ac)$ exist. Let $d = \gcd(a, b)$. Then there exist integers x and y such that $d = ax + by = ax + (-b)(-y)$. Thus, any common divisor of a and $-b$ is also a divisor of d. Hence, $\gcd(a, -b)|d$. Similarly, $d|\gcd(a, -b)$. Since $\gcd(a, b)$ and $\gcd(a, -b)$ are positive, $\gcd(a, b) = \gcd(a, -b)$.

Let $e = \gcd(a, b+ac)$. Then there exist integers p and q such that $e = ap+(b+ac)q = ap+bq+acq = a(p + cq) + bq$. Since $d|a$ and $d|b$, $d|e$. Also, $d = ax + by = ax + (b + ac)y - acy = a(x - cy) + (b + ac)y$. Since $e|a$ and $e|b + ac$, $e|d$. Hence, $e = d$.

Exercise 5 Find integers x and y such that $512x + 320y = 64$.

Solution

$$\begin{aligned} 512 &= 320 \cdot 1 + 192 \\ 320 &= 192 \cdot 1 + 128 \\ 192 &= 128 \cdot 1 + 64 \\ 128 &= 64 \cdot 2 + 0. \end{aligned}$$

Thus, $64 = 192 - 128 = 192 - (320 - 192) = 192 \cdot 2 + 320 \cdot (-1) = (512 - 320) \cdot 2 + 320 \cdot (-1) = 512 \cdot 2 + 320 \cdot (-3)$. Hence, $x = 2$ and $y = -3$.

Exercises

1. Determine $\gcd(90, 252)$. Find integers s and t such that
$$\gcd(90, 252) = s \cdot 90 + t \cdot 252.$$
2. Find integers s and t such that $\gcd(963, 652) = s \cdot 963 + t \cdot 652$.
3. Find integers s and t such that $657s + 963t = 9$.

4. Use the principle of mathematical induction to prove the following.
   (i) $1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$, $n = 1, 2, \ldots$.
   (ii) $7^n - 1$ is divisible by 6 for all $n \in \mathbb{Z}^{\#}$.
   (iii) $6 \cdot 7^n - 2 \cdot 3^n$ is divisible by 4 for all $n \in \mathbb{Z}^{\#}$.
   (iv) $5^{2n} + 3$ is divisible by 4 for all $n \in \mathbb{Z}^{\#}$.
   (v) $n < 2^n$ for all $n \in \mathbb{Z}^{\#}$.
   (vi) $2^n \geq n^2$, $n = 4, 5, \ldots$.
   (vii) $n! \geq 3^n$, $n = 7, 8, \ldots$.
5. Let a, b, and c be three integers such that $a \neq 0$. Prove the following:
   (i) If $a|b$, then $a|bc$ for all $c \in \mathbb{Z}$.
   (ii) If $b \neq 0$, $a|b$ and $b|c$, then $a|c$.
   (iii) If $a|b$ and $a|c$, then $a|(bx + cy)$ for all $x, y \in \mathbb{Z}$.
   (iv) If a, b are positive integers such that $a|b$, then $a \leq b$.
   (v) If $b \neq 0$, $a|b$, and $b|a$, then $a = \pm b$.
6. Let a, b, and c be integers. Prove that if $ac \neq 0$ and $ac|bc$, then $a|b$.
7. Let a, b, c, and d be integers such that $a \neq 0$ and $b \neq 0$. Prove that if $a|c$ and $b|d$, then $ab|cd$.
8. Let p be a prime integer, m, n integers and r a positive integer. Suppose $p^r|mn$ and $p \not| m$. Show that $p^r|n$.
9. Let a and b be integers and $\gcd(a, b) = d$. If $a = dm$ and $b = dn$, prove that $\gcd(m, n) = 1$.
10. Let a, b, and c be positive integers. Prove that $\gcd(ab, ac) = a \gcd(b, c)$.
11. Prove that if $\gcd(x, y) = \gcd(x, z) = 1$, then $\gcd(x, yz) = 1$ for all $x, y, z \in \mathbb{N}$.
12. Prove that if $\gcd(x, y) = 1, x|z$, and $y|z$, then $xy|z$ for all $x, y, z \in \mathbb{N}$.
13. Let $a, b \in \mathbb{N}$. Show that $\gcd(a, b) = \gcd(a, a + b)$.
14. Prove that $\gcd(a, b) = 1$ for any two positive consecutive integers a and b.
15. Let x and y be nonzero integers. The **least common multiple** of x and y, written $\operatorname{lcm}(x, y)$, is defined to be a positive integer m such that
    (i) $x|m$ and $y|m$ and
    (ii) if $x|c$ and $y|c$, then $m|c$.
    Prove that $\operatorname{lcm}(x, y)$ exists and is unique.
16. Let x and y be nonzero integers. Prove that $\operatorname{lcm}(x, y) \cdot \gcd(x, y) = |xy|$.
17. Let x and y be nonzero integers. Show that $\operatorname{lcm}(x, y) = |xy|$ if and only if $\gcd(x, y) = 1$.
18. Show that there are infinitely many prime integers of the form $6n - 1$, $n \geq 1$.
19. Let S be a set with n elements, $n \geq 1$. Show by mathematical induction that $|\mathcal{P}(S)| = 2^n$.
20. Determine whether the following assertions are true or false. If true, then prove it, and if false give a counterexample.
    (i) If p is a prime such that $p|a^5$, then $p|a$, where a is an integer.
    (ii) If p is a prime such that $p|(a^2 + b^2)$ and $p|a$, then $p|b$, where a and b are integers.
    (iii) For any integer a, $\gcd(a, a + 3) = 1$ or 3.
    (iv) If $\gcd(a, 6) = 3$ and $\gcd(b, 6) = 3$, then $\gcd(a + b, 6) = 6$, where a and b are integers.
    (v) If $\gcd(b, c) = 1$ and $a|b$, then $\gcd(a, c) = 1$.

1.3 Relations

Some describe or define mathematics as the study of relations. Since a relation is a set of ordered pairs, we get our first glimpse of the fundamental importance of the concept of an ordered pair.

Definition 1.3.1 *A **binary relation** or simply a **relation** R from a set A into a set B is a subset of $A \times B$.*

Let R be a relation from a set A into a set B. If $(x, y) \in R$, we write xRy or $R(x) = y$. If xRy, then sometimes we say that x is related to y (or y is in relation with x) with respect to R or simply x is related to y. If $A = B$, then we speak of a binary relation on A.

Example 1.3.2 *Let A denote the names of all states in the USA and $B = \mathbb{Z}$. With each state a in A associate an integer n which denotes the number of people in that state in the year* 1996. *Then $R = \{(a, n) \mid a \in A$ and n is the number of people in state a in* 1996$\}$ *is a subset of $A \times \mathbb{Z}$. Thus, R defines a relation from A into $\mathbb{Z}$.*

Example 1.3.3 *Consider the set of integers $\mathbb{Z}$. Let R be the set of all ordered pairs (m, n) of integers such that $m < n$, i.e.,*

$$R = \{(m, n) \in \mathbb{Z} \times \mathbb{Z} \mid m < n\}.$$

Then R is a binary relation on $\mathbb{Z}$.

Let R be a relation from a set A into a set B. By looking at the elements of R, we can find out which elements of A are related to elements of B with respect to R. The elements of A that are related to elements of B form a subset of A, called the **domain** of R, and the elements of B that are in relation with elements of A form a subset of B, called the **range** of R. More formally, we have the following definition.

Definition 1.3.4 *Let R be a relation from a set A into a set B. Then the **domain** of R, denoted by $\mathcal{D}(R)$, is defined to be the set*

$$\{x \mid x \in A \text{ and there exists } y \in B \text{ such that } (x, y) \in R\}.$$

*The **range** or **image** of R, denoted by $\mathcal{I}(R)$, is defined to be the set*

$$\{y \mid y \in B \text{ and there exists } x \in A \text{ such that } (x, y) \in R\}.$$

Example 1.3.5 *Let $A = \{4, 5, 7, 8, 9\}$ and $B = \{16, 18, 20, 22\}$. Define $R \subseteq A \times B$ by*

$$R = \{(4, 16), (4, 20), (5, 20), (8, 16), (9, 18)\}.$$

Then R is a relation from A into B. Here $(a, b) \in R$ if and only if a divides b, where $a \in A$ and $b \in B$. Note that for the domain of R, we have $\mathcal{D}(R) = \{4, 5, 8, 9\}$ and for the range of R, we have $\mathcal{I}(R) = \{16, 18, 20\}$.

Example 1.3.6 *Let $S = \{(x, y) \mid x, y \in \mathbb{R}, x^2 + y^2 = 1, y > 0\}$. Then S is a binary relation on $\mathbb{R}$. S is the set of points in the Euclidean plane constituting the semicircle lying above the x-axis with center $(0, 0)$ and radius* 1.

Definition 1.3.7 *Let R be a binary relation on a set A. Then R is called*
*(i) **reflexive** if for all $x \in A$, xRx,*
*(ii) **symmetric** if for all $x, y \in A$, xRy implies yRx,*
*(iii) **transitive** if for all $x, y, z \in A$, xRy and yRz imply xRz.*

Definition 1.3.8 *A binary relation E on a set A is called an **equivalence relation** on A if E is reflexive, symmetric, and transitive.*

The important concept of an equivalence relation is due to Gauss. We will use this concept repeatedly throughout the text.

Example 1.3.9 *Let $A = \{1, 2, 3, 4, 5, 6\}$ and $E = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), (2, 3), (3, 2)\}$. Then E is an equivalence relation on A.*

Example 1.3.10 *(i) Let L denote the set of all straight lines in the Euclidean plane and E be the relation on L defined by for all $l_1, l_2 \in L, (l_1, l_2) \in E$ if and only if l_1 and l_2 are parallel. Then E is an equivalence relation on L.*

(ii) Let L be defined as in (i) and P be the relation defined on L by for all $l_1, l_2 \in L$, $(l_1, l_2) \in P$ if and only if l_1 and l_2 are perpendicular. Let l be a line in L. Since l cannot be perpendicular to itself, $(l, l) \notin P$. Hence, P is not reflexive and so P is not an equivalence relation on L. Also, P is not transitive.

Example 1.3.11 *Let n be a fixed positive integer in $\mathbb{Z}$. Define the relation $\equiv_n$ on $\mathbb{Z}$ by for all $x, y \in \mathbb{Z}$, $x \equiv_n y$ if and only if $n|(x - y)$, i.e., $x - y = nk$ for some $k \in \mathbb{Z}$. We now show that $\equiv_n$ is an equivalence relation on $\mathbb{Z}$.*

(i) For all $x \in \mathbb{Z}$, $x - x = 0 = 0n$. Hence, for all $x \in \mathbb{Z}$, $x \equiv_n x$. Thus, $\equiv_n$ is reflexive.

(ii) Let $x, y \in \mathbb{Z}$. Suppose $x \equiv_n y$. Then there exists $q \in \mathbb{Z}$ such that $qn = x - y$. Thus, $(-q)n = y - x$ and so $n|(y - x)$, i.e., $y \equiv_n x$. Hence, $\equiv_n$ is symmetric.

(iii) Let $x, y, z \in \mathbb{Z}$. Suppose $x \equiv_n y$ and $y \equiv_n z$. Then there exist q, $r \in \mathbb{Z}$ such that $qn = x - y$ and $rn = y - z$. Thus, $(q + r)n = x - z$ and $q + r \in \mathbb{Z}$. This implies that $x \equiv_n z$. Hence, $\equiv_n$ is transitive.

Consequently, $\equiv_n$ is an equivalence relation on $\mathbb{Z}$.

The equivalence relation, $\equiv_n$, as defined in Example 1.3.11 is called **congruence modulo** n. (Another commonly used notation for $x \equiv_n y$ is $x \equiv y (\text{mod } n)$.)

Definition 1.3.12 *Let E be an equivalence relation on a set A. For all $x \in A$, let $[x]$ denote the set*

$$[x] = \{y \in A \mid yEx\}.$$

The set $[x]$ is called the **equivalence class (**with respect to E) determined by x.

In the following theorem, we prove some basic properties of equivalence classes.

Theorem 1.3.13 *Let E be an equivalence relation on the set A. Then*

(i) for all $x \in A$, $[x] \neq \emptyset$,

(ii) if $y \in [x]$, then $[x] = [y]$, where $x, y \in A$,

(iii) for all $x, y \in A$, either $[x] = [y]$ or $[x] \cap [y] = \emptyset$,

(iv) $A = \cup_{x \in A}[x]$, i.e., A is the union of all equivalence classes with respect to E.

Proof. (i) Let $x \in A$. Since E is reflexive, xEx. Hence, $x \in [x]$ and so $[x] \neq \emptyset$.

(ii) Let $y \in [x]$. Then yEx and by the symmetric property of E, xEy. In order to show that $[x] = [y]$, we will show that $[x] \subseteq [y]$ and $[y] \subseteq [x]$. The result then will follow by Theorem 1.1.1. Let $u \in [y]$. Then uEy. Since uEy and yEx, the transitivity of E implies that uEx. Hence, $u \in [x]$. Thus, $[y] \subseteq [x]$. Now let $u \in [x]$. Then uEx. Since uEx and xEy, uEy by transitivity and so $u \in [y]$. Hence, $[x] \subseteq [y]$. Consequently, $[x] = [y]$.

(iii) Let $x, y \in A$. Suppose $[x] \cap [y] \neq \emptyset$. Then there exists $u \in [x] \cap [y]$. Thus, $u \in [x]$ and $u \in [y]$, i.e., uEx and uEy. Since E is symmetric and uEy, we have yEu. Now yEu and uEx and so by the transitivity of E, yEx. This implies that $y \in [x]$. Hence, by (ii), $[y] = [x]$.

(iv) Let $x \in A$. Then $x \in [x] \subseteq \cup_{x \in A}[x]$. Thus, $A \subseteq \cup_{x \in A}[x]$. Also, $\cup_{x \in A}[x] \subseteq A$. Hence, $A = \cup_{x \in A}[x]$.

■

One of the main objectives of this section is to study the relationship between an equivalence relation and a partition of a set. We now focus our attention to partitions. We begin with the following definition.

Definition 1.3.14 *Let A be a set and $\mathcal{P}$ be a collection of nonempty subsets of A. Then $\mathcal{P}$ is called a* ***partition*** *of A if the following properties are satisfied:*

(i) for all B, $C \in \mathcal{P}$, either $B = C$ or $B \cap C = \emptyset$.

(ii) $A = \cup_{B \in \mathcal{P}} B$.

In other words, if $\mathcal{P}$ is a partition of A, then (i) $B \subseteq A$ for all $B \in \mathcal{P}$, i.e., every element of $\mathcal{P}$ is a subset of A, (ii) distinct elements of $\mathcal{P}$ are either equal or disjoint, and (iii) the union of the members of $\mathcal{P}$ is A.

Example 1.3.15 *(i) Let $A = \{1, 2, 3, 4, 5, 6\}$. Let $A_1 = \{1\}$, $A_2 = \{2, 4, 6\}$, and $A_3 = \{3, 5\}$. Now $A = A_1 \cup A_2 \cup A_3$, $A_1 \cap A_2 = \emptyset$, $A_1 \cap A_3 = \emptyset$, and $A_2 \cap A_3 = \emptyset$. Hence, $\mathcal{P} = \{A_1, A_2, A_3\}$ is a partition of A.*

(ii) Consider $\mathbb{Z}$. Let A be the set of all even integers and B be the set of all odd integers. Then $A \cap B = \emptyset$ and $A \cup B = \mathbb{Z}$. Thus, $\{A, B\}$ is a partition of $\mathbb{Z}$.

The following theorem is immediate from Theorem 1.3.13.

Theorem 1.3.16 *Let E be an equivalence relation on the set A. Then*

$$\mathcal{P} = \{[x] \mid x \in A\}$$

is a partition of A. ■

Example 1.3.17 *Consider the equivalence relation $\equiv_n$ on $\mathbb{Z}$ as defined in Example* 1.3.11. *Let $\mathbb{Z}_n = \{[x] \mid x \in \mathbb{Z}\}$. By Theorem* 1.3.16, *$\mathbb{Z}_n$ is a partition of $\mathbb{Z}$. Suppose $n = 6$. We claim that*

$$\mathbb{Z}_6 = \{[0], [1], [2], [3], [4], [5]\}$$

and

$$[i] = \{0 + i, \pm 6 + i, \pm 12 + i, \ldots\} = \{6q + i \mid q \in \mathbb{Z}\} \text{ for all } i \in \mathbb{Z}.$$

Let $0 \leq n < m < 6$. Suppose $[n] = [m]$. Then $m \in [n]$ and so $6|(m - n)$. This is a contradiction since $0 < m - n < 6$. Hence, the equivalence classes $[0], [1], [2], [3], [4], [5]$ are distinct. We now show that these are the only distinct equivalence classes.

Let k be any integer. By the division algorithm, $k = 6q + r$ for some integers q and r such that $0 \leq r < 6$. Thus, $k - r = 6q$ and so $6|(k - r)$. This implies that $k \equiv_6 r$ and so $[k] = [r]$. Since $0 \leq r < 6$ we have $[r] \in \{[0], [1], [2], [3], [4], [5]\}$ and so $[k] \in \{[0], [1], [2], [3], [4], [5]\}$. This proves our first claim.

Let $i \in \mathbb{Z}$. Then $x \in [i]$ if and only if $6|(x - i)$ if and only if $6q = x - i$ for some $q \in \mathbb{Z}$ if and only if $x = 6q + i$ for some $q \in \mathbb{Z}$. This proves our second claim. It now follows that for all $i = 0, 1, \ldots, 5$, $[i] = [6q + i]$ for all $q \in \mathbb{Z}$. Hence,

$$\begin{array}{l}
\text{for } i = 0,\ [0] = [6] = [12] = \cdots = [-6] = [-12] = \cdots; \\
\text{for } i = 1,\ [1] = [7] = [13] = \cdots = [-5] = [-11] = \cdots; \\
\text{for } i = 2,\ [2] = [8] = [14] = \cdots = [-4] = [-10] = \cdots; \\
\text{for } i = 3,\ [3] = [9] = [15] = \cdots = [-3] = [-9] = \cdots; \\
\text{for } i = 4,\ [4] = [10] = [16] = \cdots = [-2] = [-8] = \cdots; \\
\text{for } i = 5,\ [5] = [11] = [17] = \cdots = [-1] = [-7] = \cdots.
\end{array}$$

By Theorem 1.3.16, given an equivalence relation E on a set A, the set of all equivalence classes forms a partition of A. We now prove that corresponding to any partition, we can associate an equivalence relation.

Theorem 1.3.18 *Let $\mathcal{P}$ be a partition of the set A. Define a relation E on A by for all $x, y \in A$, xEy if there exists $B \in \mathcal{P}$ such that x, $y \in B$. Then E is an equivalence relation on A and the equivalence classes are precisely the elements of $\mathcal{P}$.*

Proof. Note that if two elements x and y of A are related, i.e., xEy, then x and y must belong to the same member of $\mathcal{P}$. Also, if $B \in \mathcal{P}$, then any two elements of B are related, i.e., xEy for all $x, y \in B$. We now prove the result.

Since $\mathcal{P}$ is a partition of A, $A = \cup_{B\in\mathcal{P}} B$. First we show that E is reflexive. Let x be any element of A. Then there exists $B \in \mathcal{P}$ such that $x \in B$. Since $x, x \in B$, we have xEx. Hence, E is reflexive. We now show that E is symmetric. Let xEy. Then $x, y \in B$ for some $B \in \mathcal{P}$. Thus, y, $x \in B$ and so yEx. Hence, E is symmetric. We now establish the transitivity of E. Let $x, y, z \in A$. Suppose xEy and yEz. Then x, $y \in B$ and $y, z \in C$ for some B, $C \in \mathcal{P}$. Since $y \in B \cap C$, $B \cap C \neq \emptyset$. Also, since $\mathcal{P}$ is a partition and $B \cap C \neq \emptyset$, we have $B = C$ so that x, $z \in B$. Hence, xEz. This shows that E is transitive. Consequently, E is an equivalence relation.

We now show that the equivalence classes determined by E are precisely the elements of $\mathcal{P}$. Let $x \in A$. Consider the equivalence class $[x]$. Since $A = \cup_{B\in P} B$, there exists $B \in \mathcal{P}$ such that $x \in B$. We claim that $[x] = B$. Let $u \in [x]$. Then uEx and so $u \in B$ since $x \in B$. Thus, $[x] \subseteq B$. Also, since $x \in B$, we have yEx for all $y \in B$ and so $y \in [x]$ for all $y \in B$. This implies that $B \subseteq [x]$. Hence, $[x] = B$. Finally, note that if $C \in \mathcal{P}$, then $C = [u]$ for all $u \in C$. Thus, the equivalence classes are precisely the elements of $\mathcal{P}$. ■

The relation E in Theorem 1.3.18 is called the **equivalence relation** on A **induced by the partition** $\mathcal{P}$.

New relations can be constructed from existing relations. For example, given relations R and S from a set A into a set B, we can form relations $R \cap S$, $R \cup S$, $R \backslash S$, $(A \times B) \backslash R$ in a natural way. In all these relations, the domain and range of the relations under consideration are subsets of A and B, respectively. Now given a relation R from a set A into a set B and a relation S from B into a set C, there is a relation from A into C that arises in a natural way as follows: Let us denote the new relation by T. Suppose $(a, b) \in R$ and $(b, c) \in S$. Then we make $(a, c) \in T$. Every element of T is constructed in this way. That is, $(a, c) \in T$ for some $a \in A$ and $c \in C$ if and only if there exists $b \in B$ such that $(a, b) \in R$ and $(b, c) \in S$. This relation T is called the **composition** of R and S and is denoted by $S \circ R$. Note that to form the composition of R and S, we must have the domain of S and the range of R to be subsets of the same set. More formally we have the following definition.

Definition 1.3.19 *Let R be a relation from a set A into a set B and S be a relation from B into a set C. The **composition** of R and S, denoted by $S \circ R$, is the relation from A into C defined by*

$$x(S \circ R)y \text{ if there exists } z \in B \text{ such that } xRz \text{ and } zSy$$

for all $x \in A, y \in C$.

Let R be a relation on a set A. Recursively, we define a relation R^n, $n \in \mathbb{N}$, as follows:

$$\begin{aligned} R^1 &= R \\ R^n &= R \circ R^{n-1} \text{ if } n > 1. \end{aligned}$$

Definition 1.3.20 *Let R be a relation from a set A into a set B. The **inverse** of R, denoted by R^{-1}, is the relation from B into A defined by*

$$xR^{-1}y \text{ if } yRx$$

for all $x \in B$, $y \in A$.

The following theorem gives a necessary and sufficient condition for a binary relation to be an equivalence relation.

Theorem 1.3.21 *Let R be a relation on a set A. Then R is an equivalence relation on A if and only if*
(i) $\triangle \subseteq R$, where $\triangle = \{(x, x) \mid x \in A\}$,
(ii) $R = R^{-1}$, and
(iii) $R \circ R \subseteq R$.

Proof. Suppose R is an equivalence relation. Let $(x, x) \in \triangle$, where $x \in A$. Since R is reflexive, $(x, x) \in R$. Hence, $\triangle \subseteq R$, i.e., (i) holds. Let $(x, y) \in R$. Since R is symmetric, $(y, x) \in R$. Thus, by the definition of R^{-1}, $(x, y) \in R^{-1}$. Hence, $R \subseteq R^{-1}$. On the other hand, let $(x, y) \in R^{-1}$. Then $(y, x) \in R$. Therefore, by the symmetric property, $(x, y) \in R$. Hence, $R^{-1} \subseteq R$. Thus, $R = R^{-1}$, i.e., (ii) holds. We now prove (iii). Let $(x, y) \in R \circ R$. Then there exists $z \in A$ such that $(x, z) \in R$ and $(z, y) \in R$. Since R is transitive, $(x, y) \in R$. Thus, $R \circ R \subseteq R$, i.e., (iii) holds.

Conversely, suppose that (i), (ii), and (iii) hold for R. For all $x \in A$, $(x, x) \in \Delta \subseteq R$. Thus, R is reflexive. Next, we show that R is symmetric. Let $(x, y) \in R$. Then by (ii), $(x, y) \in R^{-1}$. This implies that $(y, x) \in R$. Hence, R is symmetric. For the transitivity of R, let $(x, z) \in R$ and $(z, y) \in R$. Then $(x, y) \in R \circ R$ by the definition of composition of relations. Since $R \circ R \subseteq R$, $(x, y) \in R$. Hence, R is transitive. Consequently, R is an equivalence relation. ■

Worked-Out Exercises

Exercise 1 In $\mathbb{Z}_{10}$, which of the following equivalence classes are equal: [2], [−5], [5], [−8], [12], [15], [−3], [7], [22]?

Solution We note that $[2] = [2 + 10] = [12]$, $[-8] = [-8 + 10] = [2]$, $[12] = [12 + 10] = [22]$, $[-5] = [-5 + 10] = [5] = [5 + 10] = [15]$ and $[-3] = [-3 + 10] = [7]$. Also, $[2] \neq [5]$, $[2] \neq [7]$ and $[5] \neq [7]$. Hence, it now follows that $[2] = [12] = [-8] = [22]$, $[-5] = [5] = [15]$ and $[-3] = [7]$.

Exercise 2 Let R be a reflexive and transitive relation on a set S. Prove that $R \cap R^{-1}$ is an equivalence relation.

Solution Since $(x, x) \in R$ for all $x \in S$, $(x, x) \in R^{-1}$ for all $x \in S$. Thus, $(x, x) \in R \cap R^{-1}$ for all $x \in S$. Hence, $R \cap R^{-1}$ is reflexive. Let $(x, y) \in R \cap R^{-1}$. Then $(x, y) \in R$ and $(x, y) \in R^{-1}$. Thus, $(y, x) \in R^{-1}$ and $(y, x) \in R$. Therefore, $(y, x) \in R \cap R^{-1}$. Hence, $R \cap R^{-1}$ is symmetric. Now suppose that $(x, y), (y, z) \in R \cap R^{-1}$. Then $(x, y), (y, z) \in R$ and $(x, y), (y, z) \in R^{-1}$. Since R is transitive, $(x, z) \in R$. Now since $(x, y), (y, z) \in R^{-1}$, $(y, x), (z, y) \in R$. Since R is transitive, $(z, x) \in R$ and so $(x, z) \in R^{-1}$. Thus, $(x, z) \in R \cap R^{-1}$. Hence, $R \cap R^{-1}$ is transitive. We have thus proved that $R \cap R^{-1}$ is reflexive, symmetric, and transitive and hence $R \cap R^{-1}$ is an equivalence relation.

Exercise 3 Give an example of an equivalence relation on the set $S = \{1, 2, 3, 4, 5, 6, 7, 8\}$ such that R has exactly four equivalence classes.

Solution $R = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), (7, 7), (8, 8), (1, 2), (2, 1), (3, 4), (4, 3), (5, 6), (6, 5), (7, 8), (8, 7)\}$. The equivalence classes are $[1] = [2]$, $[3] = [4]$, $[5] = [6]$, and $[7] = [8]$.

Exercise 4 Let R_1 and R_2 be two symmetric relations on a set S. Prove that $R_1 \circ R_2$ is symmetric if and only if $R_1 \circ R_2 = R_2 \circ R_1$.

Solution Suppose $R_1 \circ R_2$ is symmetric. Let (x, y) be any element of $R_1 \circ R_2$. Then $(y, x) \in R_1 \circ R_2$ since $R_1 \circ R_2$ is symmetric. Thus, there exists $z \in S$ such that $(y, z) \in R_2$ and $(z, x) \in R_1$ by the definition of composition of relations. Since R_1 and R_2 are symmetric, $(z, y) \in R_2$ and $(x, z) \in R_1$. Hence, $(x, y) \in R_2 \circ R_1$. Thus, $R_1 \circ R_2 \subseteq R_2 \circ R_1$. Similarly, $R_2 \circ R_1 \subseteq R_1 \circ R_2$. Hence, $R_1 \circ R_2 = R_2 \circ R_1$.

Conversely, suppose that $R_1 \circ R_2 = R_2 \circ R_1$. Let $(x, y) \in R_1 \circ R_2$. Then $(x, y) \in R_2 \circ R_1$. Thus, there exists $z \in S$ such that $(x, z) \in R_1$ and $(z, y) \in R_2$. Since R_1 and R_2 are symmetric, $(z, x) \in R_1$ and $(y, z) \in R_2$. Hence, $(y, x) \in R_2 \circ R_1 = R_1 \circ R_2$. Thus, $R_1 \circ R_2$ is symmetric.

Exercise 5 Let $A = \{1, 2, 3, 4, 5\}$ and $R = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (1, 2), (2, 1), (4, 5), (5, 4)\}$. Show that R is an equivalence relation.

Solution Let $B = \{1, 2\}$, $C = \{3\}$, and $D = \{4, 5\}$. Let $\mathcal{P} = \{B, C, D\}$. Then $\mathcal{P}$ is a partition of A. Also, note that if $x, y \in A$, then $(x, y) \in R$ if and only if $x, y \in X$ for some $X \in \mathcal{P}$, i.e., the relation R is induced by the partition $\mathcal{P}$. Hence, R is an equivalence relation on A by Theorem 1.3.18.

Exercise 6 Let $X = \{1, 2, 3, 4, 5, 6, 7\}$. Then

$$\mathcal{P} = \{\{1, 3, 5\}, \{2, 6\}, \{4, 7\}\}$$

is a partition of X. List the elements of the corresponding equivalence relation R on X induced by $\mathcal{P}$.

Solution $R = \{(a, b) \in X \times X \mid a \text{ and } b \text{ both belong to the same element of } \mathcal{P}\}$. Then $R = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), (7, 7), (1, 3), (3, 1), (1, 5), (5, 1), (3, 5), (5, 3), (2, 6), (6, 2), (4, 7), (7, 4)\}$.

Exercise 7 Let R be a relation on a set S. Prove that the following conditions are equivalent.

(i) R is an equivalence relation on S.

(ii) R is reflexive and for all $a, b, c \in S$, if aRb and bRc, then cRa.

Solution (i)⇒(ii): Suppose R is an equivalence relation on S. Then R is reflexive. Let $a, b, c \in S$. Suppose aRb and bRc. The transitive property of R implies that aRc. Hence, cRa since R is symmetric.

(ii)⇒(i): Since R is given to be reflexive, to show that R is an equivalence relation, we only need to check that R is symmetric and transitive. For symmetry, suppose aRb. Since R is reflexive, we have aRa. Now since we have aRa and aRb, bRa by hypothesis. This shows that R is symmetric. To show that R is transitive, suppose aRb and bRc. Then by the hypothesis, cRa. Since we have shown that R is symmetric, cRa implies that aRc. Hence, R is transitive. Consequently, R is an equivalence relation on S.

Exercises

1. Let R be a relation on the set $A = \{1, 2, 3, 4, 5, 6, 7\}$ defined by $R = \{(a, b) \in A \times A \mid 4 \text{ divides } a - b\}$.
   (i) List the elements of R.
   (ii) Find the domain of R.
   (iii) Find the range of R.
   (iv) Find the elements of R^{-1}.
   (v) Find the domain of R^{-1}.
   (vi) Find the range of R^{-1}.
2. Let R be a relation on the set $A = \{1, 2, 3, 4, 5, 6\}$ defined by $R = \{(a, b) \in A \times A \mid a + b \leq 9\}$.
   (i) List the elements of R.
   (ii) Is $\triangle \subseteq R$, where $\Delta = \{(x, x) \mid x \in A\}$?
   (iii) Is $R = R^{-1}$?
   (iv) Is $R \circ R \subseteq R$?
3. Which of the following relations E are equivalence relations on the set of integers $\mathbb{Z}$?
   (i) xEy if and only if $x - y$ is an even integer.
   (ii) xEy if and only if $x - y$ is an odd integer.
   (iii) xEy if and only if $x \leq y$.
   (iv) xEy if and only if x divides y.
   (v) xEy if and only if $x^2 = y^2$.
   (vi) xEy if and only if $|x| = |y|$.
   (vii) xEy if and only if $|x - y| \leq 2$.
4. Let $R = \{(a, b) \mid a, b \in \mathbb{Q} \text{ and } a - b \in \mathbb{Z}\}$. Prove that R is an equivalence relation on $\mathbb{Q}$.

5. Let $A = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Define a relation R on A by

   $$aRb \text{ if and only if 3 divides } a - b$$

   for all $a, b \in A$. Show that R is an equivalence relation on A. Find the equivalence classes [1], [2], [3], and [4].

6. Let R be an equivalence relation on a set A. Find the domain and range of R.

7. Find all equivalence relations on the set $S = \{a, b, c\}$.

8. In $\mathbb{Z}_6$, which of the following equivalence classes are equal: $[-1]$, $[2]$, $[8]$, $[5]$, $[-2]$, $[11]$, $[23]$?

9. Let x, $y \in \mathbb{Z}$ be such that $x \equiv_n y$, where $n \in \mathbb{N}$. Show that for all $z \in \mathbb{Z}$,

   (i) $x + z \equiv_n y + z$,

   (ii) $xz \equiv_n yz$

10. Let $x, y, z, w \in \mathbb{Z}$ and n be a positive integer. Suppose that $x \equiv_n y$ and $z \equiv_n w$. Show that $x + z \equiv_n y + w$ and $xz \equiv_n yw$.

11. Let n be a positive integer and $[x], [y] \in \mathbb{Z}_n$. Show that the following conditions are equivalent.

    (i) $[x] = [y]$.

    (ii) $x - y = nr$ for some integer r.

    (iii) $n|(x - y)$.

12. (**Chinese Remainder Theorem**) Let m and n be positive integers such that $\gcd(m, n) = 1$. Prove that for any integers a and b, the congruences $x \equiv_m a$ and $x \equiv_n b$ have a common solution in $\mathbb{Z}$. Furthermore, if u and v are two solutions of these congruences, prove that $u \equiv_{mn} v$.

13. Define relations R_1, R_2, R_3 such that R_1 is reflexive and symmetric but not transitive, R_2 is reflexive and transitive but not symmetric, and R_3 is symmetric and transitive but not reflexive.

14. Prove that the intersection of two equivalence relations on a set S is an equivalence relation on S.

15. Let R be a relation on a set A. Define $\mathcal{T}(R) = R \cup R^{-1} \cup \{(x, x) \mid x \in A\}$. Show that $\mathcal{T}(R)$ is reflexive and symmetric.

16. Let R be a relation on a set S. Set $R^\infty = R \cup R^2 \cup R^3 \cup \cdots$. Prove the following:

    (i) R^∞ is a transitive relation on S.

    (ii) If T is a transitive relation on A such that $R \subseteq T$, then $R^\infty \subseteq T$.

    (R^∞ is called the **transitive closure** of R.)

17. Let R_1 and R_2 be symmetric relations on a set S such that $R_1 \circ R_2 \subseteq R_2 \circ R_1$. Prove that $R_2 \circ R_1$ is symmetric and $R_1 \circ R_2 = R_2 \circ R_1$.

18. Let R_1 and R_2 be equivalence relations on a set S such that $R_1 \circ R_2 = R_2 \circ R_1$. Prove that $R_1 \circ R_2$ is an equivalence relation.

19. Let R_1 and R_2 be relations on a set S. Determine whether each statement is true or false. If the statement is false, give a counterexample.

    (i) If R_1 and R_2 are reflexive, then $R_1 \circ R_2$ is reflexive.

    (ii) If R_1 and R_2 are transitive, then $R_1 \circ R_2$ is transitive.

    (iii) If R_1 and R_2 are symmetric, then $R_1 \circ R_2$ is symmetric.

    (iv) If R_1 is transitive, then R_1^{-1} is transitive.

    (v) If R_1 is reflexive and transitive, then $R_1 \circ R_1$ is transitive.

1.4 Functions

Like sets, functions play a central role in mathematics. Readers may already be familiar with the notion of a function either through a college algebra or a calculus course. In these courses, functions were usually real valued. Throughout the text we will encounter functions which do not have to be real valued. Functions help us study the relationship between various algebraic structures. In this section, we review some of their basic properties. Roughly speaking, a function is a special type of correspondence between elements of one set and those of another set. More precisely, a function is a particular set of ordered pairs.

Definition 1.4.1 *Let A and B be nonempty sets. A relation f from A into B is called a* ***function*** *(or* ***mapping****) from A into B if*
(i) $\mathcal{D}(f) = A$ and
(ii) for all $(x, y), (x', y') \in f$, $x = x'$ implies $y = y'$.
When (ii) is satisfied by a relation f, we say that f is ***well defined*** *or* ***single-valued****.*

We use the notation $f : A \to B$ to denote a function f from a set A into a set B. For $(x, y) \in f$, we usually write $f(x) = y$ and say that y is the **image** of x under f and x is a **preimage** of y under f.

Leibniz seems to be the first to have used the word "function" to stand for any quantity related to a curve. Clairant (1734) originated the notation $f(x)$ and Euler made extensive use of it. Dirichlet is responsible for the current definition of a function.

Let us now explain the above definition. Suppose $f : A \to B$. Then f is a subset of $A \times B$ such that for all $x \in A$, there exists a unique $y \in B$ such that $(x, y) \in f$. Hence, we like to think of a function as a rule which associates to each element x of A exactly one element y of B. In order to show that a relation f from A into B is a function, we first show that the domain of f is A and next we show that f well defined or single-valued, i.e., if $x = y$ in A, then $f(x) = f(y)$ in B for all $x, y \in A$.

We now consider some examples of relations, some of which are functions and some of which are not.

Example 1.4.2 *Let f be the subset of $\mathbb{Z}\times\mathbb{Z}$ defined by*

$$f = \{(n, 2n + 3) \mid n \in \mathbb{Z}\}.$$

Then $\mathcal{D}(f) = \{n \mid n \in \mathbb{Z}\} = \mathbb{Z}$. We now show that f is well defined. Let $n, m \in \mathbb{Z}$. Suppose $n = m$. Then $2n + 3 = 2m + 3$, i.e., $f(n) = f(m)$. Therefore, f is well defined. Hence, f satisfies (i) and (ii) of Definition 1.4.1 *and so f is a function.*

Example 1.4.3 *Let $A = \{1, 2, 3, 4\}$ and $B = \{a, b, c\}$. Let f be the subset of $A \times B$ defined by*

$$f = \{(1, a), (2, b), (3, c), (4, b)\}.$$

First note that $\mathcal{D}(f) = \{1, 2, 3, 4\} = A$ and so f satisfies (i) of Definition 1.4.1*. From the definition of f, it is immediate that for all $x \in A$, there exists a unique $y \in B$ such that $(x, y) \in f$. Therefore, f is well defined and so f satisfies (ii) of Definition* 1.4.1*. Hence, f is a function.*

Example 1.4.4 *Let f be the subset of $\mathbb{Q}\times\mathbb{Z}$ defined by*

$$f = \{(\frac{p}{q}, p) \mid p, q \in \mathbb{Z},\ q \neq 0\}.$$

First we note that $\mathcal{D}(f) = \{\frac{p}{q} \mid p, q \in \mathbb{Z},\ q \neq 0\} = \mathbb{Q}$. Thus, f satisfies (i) of Definition 1.4.1*. Now $(\frac{2}{3}, 2) \in f$, $(\frac{4}{6}, 4) \in f$ and $\frac{2}{3} = \frac{4}{6}$. But $f(\frac{2}{3}) = 2 \neq 4 = f(\frac{4}{6})$. Thus, f is not well defined. Hence, f is not a function from $\mathbb{Q}$ into $\mathbb{Z}$.*

Example 1.4.5 *Let f be the subset of $\mathbb{Z}\times\mathbb{Z}$ defined by*

$$f = \{(mn, m + n) \mid m, n \in \mathbb{Z}\}.$$

First we show that f satisfies (i) of Definition 1.4.1. *Let x be any element of $\mathbb{Z}$. Then we can write $x = x \cdot 1$. Hence, $(x, x+1) = (x \cdot 1, x+1) \in f$. This implies that $x \in \mathcal{D}(f)$. Thus, $\mathbb{Z} \subseteq \mathcal{D}(f)$. However, $\mathcal{D}(f) \subseteq \mathbb{Z}$ and so $\mathcal{D}(f) = \mathbb{Z}$. Thus, f satisfies (i) of Definition* 1.4.1. *Now $4 \in \mathbb{Z}$ and $4 = 4 \cdot 1 = 2 \cdot 2$. Thus, $(4 \cdot 1, 4+1) \in f$ and $(2 \cdot 2, 2+2) \in f$. Hence, we find that $4 \cdot 1 = 2 \cdot 2$ and $f(4 \cdot 1) = 5 \neq 4 = f(2 \cdot 2)$. This implies that f is not well defined, i.e., f does not satisfy (ii) of Definition* 1.4.1. *Hence, f is not a function from $\mathbb{Z}$ into $\mathbb{Z}$.*

We now explore the meaning of equality of two functions.

Let $f : A \to B$ and $g : A \to B$ be two functions. Then f and g are subsets of $A \times B$. Suppose $f = g$. Let x be any element of A. Then $(x, f(x)) \in f = g$. Also, $(x, g(x)) \in g$. Since g is a function and $(x, f(x)), (x, g(x)) \in g$, we must have $g(x) = f(x)$. Conversely, assume that $g(x) = f(x)$ for all $x \in A$. Let $(x, y) \in f$. Then $y = f(x) = g(x)$. Thus, $(x, y) \in g$. This implies that $f \subseteq g$. Similarly, we can show that $g \subseteq f$. It now follows that $f = g$. Thus, two functions $f : A \to B$ and $g : A \to B$ are **equal** if and only if $f(x) = g(x)$ for all $x \in A$.

Example 1.4.6 *Let $f : \mathbb{Z} \to \mathbb{Z}^{\#}$ and $g : \mathbb{Z} \to \mathbb{Z}^{\#}$ be defined by $f = \{(n, n^2) \mid n \in \mathbb{Z}\}$ and $g = \{(n, |n|^2) \mid n \in \mathbb{Z}\}$. Now for all $n \in \mathbb{Z}$,*

$$f(n) = n^2 = |n|^2 = g(n).$$

Hence, $f = g$.

Definition 1.4.7 *Let f be a function from a set A into a set B. Then*

*(i) f is called **one-one** if for all $x, x' \in A$, $f(x) = f(x')$ implies $x = x'$.*

*(ii) f is called **onto** B (or f **maps** A **onto** B) if $\mathcal{I}(f) = B$.*

We note that if $f : A \to B$, then $\mathcal{I}(f) = B$ if and only if for all $y \in B$, there exists $x \in A$ such that $f(x) = y$. In other words, $\mathcal{I}(f) = B$ if and only if every element of B has a preimage. We also note that f is one-one if and only if every element of B has at most one preimage.

Let A be a nonempty set. The function $i_A : A \to A$ defined by $i_A(x) = x$ for all $x \in A$ is a one-one function of A onto A. i_A is called the **identity map** on A.

Example 1.4.8 *Consider the relation f from $\mathbb{Z}$ into $\mathbb{Z}$ defined by*

$$f(n) = n^2$$

for all $n \in \mathbb{Z}$. Now $\mathcal{D}(f) = \mathbb{Z}$. Also, if $n = n'$, then $n^2 = (n')^2$, i.e., $f(n) = f(n')$. Hence, f is well defined. Thus, f is a function. Now $f(1) = 1 = f(-1)$ and $1 \neq -1$. This implies that f is not one-one. Now for all $n \in \mathbb{Z}$, $f(n)$ is a nonnegative integer. This shows that a negative integer has no preimage. Hence, f is not onto $\mathbb{Z}$. Note that f is onto $\{0, 1, 4, 9, \ldots\}$.

Example 1.4.9 *Consider the relation f from $\mathbb{Z}$ into $\mathbb{Z}$ defined by for all $n \in \mathbb{Z}$, $f(n) = 2n$. As in the previous examples, we can show that f is a function. Let $n, n' \in \mathbb{Z}$ and suppose that $f(n) = f(n')$. Then $2n = 2n'$, i.e., $n = n'$. Hence, f is a one-one function. Since for all $n \in \mathbb{Z}$, $f(n)$ is an even integer, we see that an odd integer has no preimage. Thus, f is not onto $\mathbb{Z}$. However, we note that f is onto $\mathbb{E}$.*

Definition 1.4.10 *Let A, B, and C be nonempty sets and $f : A \to B$ and $g : B \to C$. The **composition** $\circ$ of f and g, written $g \circ f$, is the relation from A into C defined as follows:*

$$\begin{aligned} g \circ f \quad = \quad & \{(x, z) \mid x \in A, z \in C, \text{ there exists } y \in B \\ & \text{such that } f(x) = y \text{ and } g(y) = z\}. \end{aligned}$$

Let $f : A \to B$ and $g : B \to C$ and $(x, z) \in g \circ f$, i.e., $(g \circ f)(x) = z$. Then by the definition of composition of functions, there exists $y \in B$ such that $f(x) = y$ and $g(y) = z$. Now

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

Hence, $(g \circ f)(x) = g(f(x))$.

In the following, we describe some properties of composition of functions.

Theorem 1.4.11 *Suppose that $f : A \to B$ and $g : B \to C$. Then*
(i) $g \circ f : A \to C$, i.e., $g \circ f$ is a function from A into C.
(ii) If f and g are one-one, then $g \circ f$ is one-one.
(iii) If f is onto B and g is onto C, then $g \circ f$ is onto C.

Proof. (i) Let $x \in A$. Since f is a function and $x \in A$, there exists $y \in B$ such that $f(x) = y$. Now since g is a function and $y \in B$, there exists $z \in C$ such that $g(y) = z$. Thus, $(g \circ f)(x) = g(f(x)) = g(y) = z$, i.e., $(x, z) \in g \circ f$. Hence, $x \in \mathcal{D}(g \circ f)$. This shows that $A \subseteq \mathcal{D}(g \circ f)$. But $\mathcal{D}(g \circ f) \subseteq A$ and so $\mathcal{D}(g \circ f) = A$. Next, we show that $g \circ f$ is well defined.

Suppose that $(x, z) \in g \circ f$, $(x_1, z_1) \in g \circ f$ and $x = x_1$, where $x, x_1 \in A$ and $z, z_1 \in C$. By the definition of composition of functions, there exist $y, y_1 \in B$ such that $f(x) = y$, $g(y) = z$, $f(x_1) = y_1$ and $g(y_1) = z_1$. Since f is a function and $x = x_1$, we have $y = y_1$. Similarly, since g is a function and $y = y_1$, we have $z = z_1$. Thus, $g \circ f$ is well defined. Hence, $g \circ f$ is a function from A into C.

(ii) Let $x, x' \in A$. Suppose $(g \circ f)(x) = (g \circ f)(x')$. Then $g(f(x)) = g(f(x'))$. Since g is one-one, $f(x) = f(x')$. Since f is one-one, $x = x'$. Thus, $g \circ f$ is one-one.

(iii) Let $z \in C$. Then there exists $y \in B$ such that $g(y) = z$ since g is onto C. Since f is onto B, there exists $x \in A$ such that $f(x) = y$. Thus, $(g \circ f)(x) = g(f(x)) = g(y) = z$. Hence, $g \circ f$ is onto C. ■

Example 1.4.12 *Consider the function $f : \mathbb{Z} \to \mathbb{Z}$ and $g : \mathbb{Z} \to \mathbb{E}$, where $f(n) = n^2$ and $g(n) = 2n$ for all $n \in \mathbb{Z}$. Then $g \circ f : \mathbb{Z} \to \mathbb{E}$ and $(g \circ f)(n) = g(f(n)) = g(n^2) = 2n^2$.*

Theorem 1.4.13 *Let $f : A \to B$, $g : B \to C$, and $h : C \to D$. Then*

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

That is, composition of functions is associative.

Proof. First note that $h \circ (g \circ f) : A \to D$ and $(h \circ g) \circ f : A \to D$. Let $x \in A$. Then

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

Thus, by the equality of two functions, $h \circ (g \circ f) = (h \circ g) \circ f$. ■

Let A be a set and $f : A \to A$. Recursively, we define

$$\begin{aligned} f^1(x) &= f(x) \\ f^{n+1}(x) &= (f \circ f^n)(x) \end{aligned}$$

for all $x \in A$, $n \in \mathbb{N}$.

Let A and B be sets. A and B are said to be **equipollent**, written $A \sim B$, if there exists a one-one function from A onto B, i.e., the elements of A and B are in **one-one correspondence.**

From Theorem 1.4.11, it follows that $\sim$ is an equivalence relation. If $A \sim B$, then sometimes we write $|A| = |B|$. It is immediate that if A and B are finite sets, then $|A| = |B|$ if and only if A and B have the same number of elements.

The following lemma, which follows from Theorem 1.4.11(ii), is of independent interest. We give a direct proof of this result.

Lemma 1.4.14 *Let A be a set and $f : A \to A$ be a one-one function. Then $f^n : A \to A$ is a one-one function for all integers $n \geq 1$.*

Proof. Suppose there exists $n > 1$ such that f^n is not one-one. Let $m > 1$ be the smallest positive integer such that f^m is not one-one. Then there exist x, $y \in A$ such that $x \neq y$ and $f^m(x) = f^m(y)$. But then $f(f^{m-1}(x)) = f(f^{m-1}(y))$ and hence $f^{m-1}(x) = f^{m-1}(y)$ since f is one-one. Now since m is the smallest positive integer such that f^m is not one-one, f^{m-1} is one-one. Hence, $x = y$, which is a contradiction. Thus, f^n is one-one for all $n \geq 1$. ■

That one-one functions on a finite set are onto is proved next.

Theorem 1.4.15 *Let A be a finite set. If $f : A \to A$ is one-one, then f is onto A.*

Proof. Let $y \in A$. Now $f^n(y) \in A$ for all $n \geq 1$. Hence,

$$\{y, f(y), f^2(y), \ldots\} \subseteq A.$$

Since A is finite, all elements of the set $\{y, f(y), f^2(y), \ldots\}$ cannot be distinct. Thus, there exist positive integers s and t such that $s > t$ and $f^s(y) = f^t(y)$. Then $f^t(f^{s-t}(y)) = f^t(y)$. Hence, $f^{s-t}(y) = y$ since by Lemma 1.4.14, f^t is one-one. Let $x = f^{s-t-1}(y) \in A$. Then $f(x) = y$. Hence, f is onto A. ■

Definition 1.4.16 *Let A and B be sets and $f : A \to B$.*
*(i) f is called **left invertible** if there exists $g : B \to A$ such that*

$$g \circ f = i_A.$$

*(ii) f is called **right invertible** if there exists $h : B \to A$ such that*

$$f \circ h = i_B.$$

A function $f : A \to B$ is called **invertible** if f is both left and right invertible.

Example 1.4.17 *Let $f : \mathbb{Z} \to \mathbb{Z}$ and $g : \mathbb{Z} \to \mathbb{Z}$ be as defined below.*

$$f(n) = 3n$$

$$g(n) = \begin{cases} \frac{n}{3} \text{ if } n \text{ is a multiple of } 3 \\ 0 \text{ if } n \text{ is not a multiple of } 3 \end{cases}$$

for all $n \in \mathbb{Z}$. Now

$$\begin{aligned} (f \circ g)(n) &= f(g(n)) \\ &= \begin{cases} n \text{ if } n \text{ is a multiple of } 3 \\ 0 \text{ if } n \text{ is not a multiple of } 3. \end{cases} \end{aligned}$$

Hence, $f \circ g \neq i_{\mathbb{Z}}$. But $(g \circ f)(n) = g(f(n)) = g(3n) = n$ for all $n \in \mathbb{Z}$. Thus, $g \circ f = i_{\mathbb{Z}}$. Hence, g is a left inverse of f.

Often we are required to find a left (right) inverse of a function. However, not every function has a left (right) inverse. Thus, before we attempt to find a left (right) inverse of a function, it would be helpful to know if a given function has a left (right) inverse or not. The following theorem is very useful in determining whether a function is left (right) invertible or invertible.

Theorem 1.4.18 *Let A and B be sets and $f : A \to B$. Then the following assertions hold.*
(i) f is one-one if and only if f is left invertible.
(ii) f is onto B if and only if f is right invertible.
(iii) f is one-one and onto B if and only if f is invertible.

Proof. (i) Suppose f is left invertible. Then there exists $g : B \to A$ such that $g \circ f = i_A$. Let x, $y \in A$ be such that $f(x) = f(y)$. Then $g(f(x)) = g(f(y))$ or $(g \circ f)(x) = (g \circ f)(y)$. Hence, $i_A(x) = i_A(y)$, i.e., $x = y$. Thus, f is one-one.

Conversely, suppose f is one-one. Then for $y \in B$, either y has no preimage or there exists a unique $x_y \in A$ such that $f(x_y) = y$. Fix $x \in A$. Define $g : B \to A$ by

$$g(y) = \begin{cases} x & \text{if } y \text{ has no preimage under } f \\ x_y & \text{if } y \text{ has a preimage under } f \text{ and } f(x_y) = y \end{cases}$$

for all $y \in B$. By the definition of g, $\mathcal{D}(g) = B$. To show g is well defined, suppose $y, y' \in B$ and $y = y'$. Then either both y and y' have no preimages or there exist unique $x_y, x_{y'} \in A$ such that $f(x_y) = y$ and $f(x_{y'}) = y'$. Suppose both y and y' have no preimages. Then $g(y) = x = g(y')$. Now suppose there exist unique $x_y, x_{y'} \in A$ such that $f(x_y) = y$ and $f(x_{y'}) = y'$. Thus, $g(y) = x_y$ and $g(y') = x'_y$. Since $y = y'$, we have $f(x_y) = f(x_{y'})$. Since f is one-one, $x_y = x_{y'}$ and so $g(y) = g(y')$. We have thus shown that g is

well defined and so g is a function. We now show that $g \circ f = i_A$. Let $u \in A$ and suppose $f(u) = v$ for some $v \in B$. Then by the definition of g, $g(v) = u$. Thus,

$$(g \circ f)(u) = g(f(u)) = g(v) = u = i_A(u).$$

Hence, $g \circ f = i_A$.

(ii) Suppose f is right invertible. Then there exists $g : B \to A$ such that $f \circ g = i_B$. Let $y \in B$. Let $x = g(y) \in A$. Now $y = i_B(y) = (f \circ g)(y) = f(g(y)) = f(x)$. Hence, f is onto B.

Conversely, suppose f is onto B. Let $y \in B$. Since f is onto, there exists $x \in A$ such that $f(x) = y$. Let $A_y = \{x \in A \mid f(x) = y\}$. Then $A_y \neq \emptyset$. Choose $x_y \in A_y$ for all $y \in B$. Define $h : B \to A$ such that $h(y) = x_y$ for all $y \in B$. Then h is a function. Let $y \in B$. Then $(f \circ h)(y) = f(h(y)) = f(x_y) = y = i_B(y)$. Hence, $f \circ h = i_B$ and so f is right invertible.

(iii) The result here follows from (i) and (ii). ■

Let $f : A \to B$ be invertible. Let g be a left inverse of f and h be a right inverse of f. Then $g \circ f = i_A$ and $f \circ h = i_B$. Now $g = g \circ i_B = g \circ (f \circ h) = (g \circ f) \circ h = i_A \circ h = h$. Thus, if f is invertible, then left and right inverses of f are the same. This also proves that the inverse of a function, if it exists, is unique.

If f is an invertible function, then the inverse of f is denoted by f^{-1}.

Let $f : A \to B$ and $A' \subseteq A$. Then f induces a function from A' into B in a natural way as defined next.

Definition 1.4.19 *Let $f : A \to B$ and A' be a nonempty subset of A. The **restriction of** f to A', written $f|_{A'}$, is defined to be*

$$f|_{A'} = \{(x', f(x')) \mid x' \in A'\}.$$

We see that $f|_{A'}$ is really the function f except that we are considering f on a smaller domain.

Definition 1.4.20 *Let $f : A' \to B$ and A be a set containing A'. A function $g : A \to B$ is called an **extension** of f to A if $g|_{A'} = f$.*

Example 1.4.21 *Consider the function $f : \mathbb{E} \to \mathbb{Z}$ and $g : \mathbb{Z} \to \mathbb{Z}$, where $f(2n) = 2n+1$ and $g(n) = n+1$ for all $n \in \mathbb{Z}$. Then g is an extension of f to $\mathbb{Z}$ and f is the restriction of g to $\mathbb{E}$. Let the function $h : \mathbb{Z} \to \mathbb{Z}$ be defined by for all $m \in \mathbb{Z}$, $h(m) = m + 1$ if $m \in \mathbb{E}$ and $h(m) = m$ if $m \notin \mathbb{E}$. Then h is an extension of f to $\mathbb{Z}$. However, $h \neq g$. Thus, a function may have more than one extension.*

In Section 1.1, we defined the Cartesian cross product, $A \times B$, of two sets A and B. We now extend this notion to a family of sets $\{A_\alpha \mid \alpha \in I\}$, where I is an index set. First let us make the following observation: Suppose $I = \{1, 2\}$. Let S be the set of all functions $f : I \to A \cup B$ such that $f(1) \in A$ and $f(2) \in B$. Then every function $f \in S$ defines an ordered pair $(f(1), f(2)) \in A \times B$. Conversely, given $x \in A$ and $y \in B$, define $f \in S$ by $f(1) = x$ and $f(2) = y$. Then the ordered pair (x, y) defines a function $f \in S$. Hence, there is a one-one correspondence between the elements of S and $A \times B$. We now define the Cartesian product of $\{A_\alpha \mid \alpha \in I\}$.

Let $\{A_\alpha \mid \alpha \in I\}$ be a family of sets. The **Cartesian (cross) product** of $\{A_\alpha \mid \alpha \in I\}$, denoted by $\prod_{\alpha \in I} A_\alpha$, is defined to be the set

$$\{f \mid f : I \to \cup_{\alpha \in I} A_\alpha \text{ and } f(\alpha) \in A_\alpha \text{ for all } \alpha \in I\}.$$

Let $f \in \prod_{\alpha \in I} A_\alpha$. Then $f(\alpha) \in A_\alpha$ for all $\alpha \in I$. Let us write $f(\alpha) = x_\alpha$ for all $\alpha \in I$. We usually write $(x_\alpha)_{\alpha \in I}$ for f, i.e., a typical member of $\prod_{\alpha \in I} A_\alpha$ is denoted by $(x_\alpha)_{\alpha \in I}$, where $x_\alpha \in A_\alpha$ for all $\alpha \in I$.

Suppose $I = \{1, 2, \ldots, n\}$ is a finite set. Then the Cartesian product $\prod_{i \in I} A_\alpha$, is denoted by $A_1 \times A_2 \times \cdots \times A_n$. A typical member of $A_1 \times A_2 \times \cdots \times A_n$ is denoted by $(x_1, x_2, \ldots, x_n)$, $x_i \in A_i$ for all $i = 1, 2, \ldots, n$. The elements of $A_1 \times A_2 \times \cdots \times A_n$ are called **ordered n-tuples**. For two elements $(x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) \in A_1 \times A_2 \times \cdots \times A_n$, $(x_1, x_2, \ldots, x_n) = (y_1, y_2, \ldots, y_n)$ if and only if $x_i = y_i$ for all i.

Worked-Out Exercises

Exercise 1 Determine which of the following mappings $f : \mathbb{R} \to \mathbb{R}$ are one-one and which are onto $\mathbb{R}$:

(i) $f(x) = x + 4$,
(ii) $f(x) = x^2$
for all $x \in \mathbb{R}$.

Solution (i) Let $x, y \in \mathbb{R}$. Suppose $f(x) = f(y)$. Then $x + 4 = y + 4$ or $x = y$. Hence, f is one-one. Now f is onto $\mathbb{R}$ if and only if for all $y \in \mathbb{R}$ there exists $x \in \mathbb{R}$ such that $f(x) = y$. Let $y \in \mathbb{R}$. If $f(x) = y$, then $x + 4 = y$ or $x = y - 4$. Also, $y - 4 \in \mathbb{R}$. Thus, we can take x to be $y - 4$. Now $f(y - 4) = y - 4 + 4 = y$. Hence, f is onto $\mathbb{R}$.

(ii) We note that $f(x)$ is a nonnegative real number for all $x \in \mathbb{R}$. This means that negative real numbers have no preimages. In particular, for all $x \in \mathbb{R}$, $f(x) = x^2 \neq -1$. Hence, f is not onto $\mathbb{R}$. Also, $f(-1) = 1 = f(1)$ and $-1 \neq 1$. Thus, f is not one-one. Thus, f is neither one-one nor onto $\mathbb{R}$.

Exercise 2 (i) Let $f : \mathbb{Z} \to \mathbb{Z}$ be a mapping defined by

$$f(x) = \begin{cases} x \text{ if } x \text{ is even} \\ 2x + 1 \text{ if } x \text{ is odd} \end{cases}$$

for all $x \in \mathbb{Z}$. Find a left inverse of f if one exists.

(ii) Let $f : \mathbb{Z} \to \mathbb{Z}$ be the mapping defined by $f(x) = |x| + x$ for all $x \in \mathbb{Z}$. Find a right inverse of f if one exists.

Solution (i) By Theorem 1.4.18, f has a left inverse if and only if f is one-one. Before we attempt to find a left inverse of f, let us first check whether f is one-one or not. Let $x, y \in \mathbb{R}$ and $f(x) = f(y)$. By the definition of f, $f(x)$ is even if x is even and $f(x)$ is odd if x is odd. Thus, since $f(x) = f(y)$, we have both x and y are either even or odd. If x and y are both even then $f(x) = x$ and $f(y) = y$ and so $x = y$. Suppose x and y are odd. Then $f(x) = 2x + 1$ and $f(y) = 2y + 1$. Then $2x + 1 = 2y + 1$ or $x = y$. Hence, f is one-one and so f has a left inverse. Thus, there exists a function $g : \mathbb{Z} \to \mathbb{Z}$ such that $g \circ f = i_{\mathbb{Z}}$. Let $x \in \mathbb{Z}$. Suppose x is even. Now $x = i_{\mathbb{Z}}(x) = (g \circ f)(x) = g(f(x)) = g(x)$. This means $g(x) = x$ when x is even. Now suppose x is odd. Then $x = i_{\mathbb{Z}}(x) = (g \circ f)(x) = g(f(x)) = g(2x + 1)$. Put $t = 2x + 1$. Then $x = \frac{t-1}{2}$. This shows that $g(x) = \frac{x-1}{2}$ if x is odd. Thus, our choice of g is

$$g(x) = \begin{cases} x & \text{if } x \text{ is even} \\ \frac{x-1}{2} & \text{if } x \text{ is odd.} \end{cases}$$

(ii) Note that $f(x) = |x| + x \geq 0$ for all $x \in \mathbb{Z}$. This shows that negative integers do not belong to $\mathcal{I}(f)$. In particular, $f(x) \neq -1$ for all $x \in \mathbb{Z}$. Thus, f is not onto $\mathbb{Z}$ and so f does not have a right inverse.

Exercise 3 Let X and Y be nonempty sets and $f : X \to Y$. If $T \subseteq X$, then $f(T)$ denotes the set $\{f(x) \mid x \in T\}$. $f(T)$ is called the **image** of T under f. Prove that f is one-one if and only if

$$f(A \cap B) = f(A) \cap f(B)$$

for all nonempty subsets A and B of X.

Solution Suppose that f is one-one. Let A and B be nonempty subsets of X. Let $y \in f(A \cap B)$. Then $y = f(x)$ for some $x \in A \cap B$. Hence, $y \in f(A) \cap f(B)$. Thus, $f(A \cap B) \subseteq f(A) \cap f(B)$. Now let $y \in f(A) \cap f(B)$. Then $y \in f(A)$ and $y \in f(B)$. Thus, $y = f(a)$ for some $a \in A$ and $y = f(b)$ for some $b \in B$. Since f is one-one and $f(a) = f(b)$, we find that $a = b$. Thus, $y \in f(A \cap B)$. Hence, $f(A) \cap f(B) \subseteq f(A \cap B)$. Consequently, $f(A \cap B) = f(A) \cap f(B)$.

Conversely, suppose that $f(A \cap B) = f(A) \cap f(B)$ for all subsets A and B of X. Suppose f is not one-one. Then there exist $x, y \in X$ such that $f(x) = f(y)$ and $x \neq y$. Let $A = \{x\}$ and $B = \{y\}$. Since $A \cap B = \emptyset$, $f(A \cap B) = \emptyset$. However, $f(A) \cap f(B) = \{f(x)\} \neq \emptyset$. Thus, $f(A \cap B) \neq f(A) \cap f(B)$, a contradiction. Hence, f is one-one.

Exercise 4 Let A be a nonempty set and E be an equivalence relation on A. Let $B = \{[x] \mid x \in A\}$, i.e., B is the set of all equivalence classes with respect to E. Prove that there exists a function f from A onto B. The set B is usually denoted by A/E and is called the **quotient set** of A determined by E.

Solution Define $f : A \to B$ by $f(x) = [x]$ for all $x \in A$. By the definition of f, $\mathcal{D}(f) = A$. Let $x, y \in A$. Suppose $x = y$. Then $[x] = [y]$ and so $f(x) = f(y)$. Thus, f is well defined. Let $[a] \in B$. Then $a \in A$ and $f(a) = [a]$. Hence, f is onto B.

Exercise 5 Let $S = \{x \in \mathbb{R} \mid -1 < x < 1\}$. Show that $\mathbb{R} \sim S$.

Solution Define $f : \mathbb{R} \to S$ by

$$f(x) = \frac{x}{1 + |x|}$$

for all $x \in \mathbb{R}$. Let $x \in \mathbb{R}$. Then $-|x| \leq x \leq |x|$, $-1 - |x| < -|x|$, and $|x| \leq 1 + |x|$. Hence, $-1 - |x| < x < 1 + |x|$. Thus, $-1 < \frac{x}{1+|x|} < 1$ and so $-1 < f(x) < 1$. This shows that $f(x) \in S$. Let $x, y \in \mathbb{R}$ and $f(x) = f(y)$. Then $\frac{x}{1+|x|} = \frac{y}{1+|y|}$. Thus, $\frac{|x|}{1+|x|} = \frac{|y|}{1+|y|}$. This implies that $|x| + |x||y| = |y| + |x||y|$ and so $|x| = |y|$. Now $\frac{x}{1+|x|} = \frac{y}{1+|y|}$ implies that $x \geq 0$ if and only if $y \geq 0$. Therefore, since $|x| = |y|$, $x = y$. Thus, f is one-one.

Now let $z \in \mathbb{R}$ and $-1 < z < 1$. If $0 \leq z < 1$, then

$$f(\frac{z}{1-z}) = \frac{\frac{z}{1-z}}{1 + \left|\frac{z}{1-z}\right|} = \frac{\frac{z}{1-z}}{1 + \frac{z}{1-z}} = z.$$

If $-1 < z < 0$, then

$$f(\frac{z}{1+z}) = \frac{\frac{z}{1+z}}{1 + \left|\frac{z}{1+z}\right|} = \frac{\frac{z}{1+z}}{1 + \frac{-z}{1+z}} = z.$$

Hence, f is onto $\mathbb{R}$. Consequently, $\mathbb{R} \sim S$.

Exercises

1. Determine which of the following mappings $f : \mathbb{R} \to \mathbb{R}$ are one-one and which are onto $\mathbb{R}$:

   (i) $f(x) = x + 1$,

   (ii) $f(x) = x^3$,

   (iii) $f(x) = |x| + x$

   for all $x \in \mathbb{R}$.

2. Consider the function $f = \{(x, x^2) \mid x \in S\}$ of $S = \{-3, -2, -1, 0, 1, 2, 3\}$ into $\mathbb{Z}$. Is f one-one? Is f onto $\mathbb{Z}$?

3. Let $f : \mathbb{R}^+ \to \mathbb{R}^+$ and $g : \mathbb{R}^+ \to \mathbb{R}^+$ be functions defined by $f(x) = \sqrt{x}$ and $g(x) = 3x + 1$ for all $x \in \mathbb{R}^+$, where $\mathbb{R}^+$ is the set of all positive real numbers. Find $f \circ g$ and $g \circ f$. Is $f \circ g = g \circ f$?

4. Let $f : \mathbb{Q}^+ \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = 1 + \frac{1}{x}$ for all $x \in \mathbb{Q}^+$ and $g(x) = x + 1$ for all $x \in \mathbb{R}$, where $\mathbb{Q}^+$ is the set of all positive rational numbers. Find $g \circ f$.

5. For each of the mappings $f : \mathbb{Z} \to \mathbb{Z}$ given below, find a left inverse of f whenever one exists.
   (i) $f(x) = x + 2$,
   (ii) $f(x) = 2x$,
   (iii) $f(x) = \begin{cases} \frac{x}{2} \text{ if } x \text{ is even} \\ 5 \text{ if } x \text{ is odd} \end{cases}$
   for all $x \in \mathbb{Z}$.

6. For each of the mappings $f : \mathbb{Z} \to \mathbb{Z}$ given below, find a right inverse of f whenever one exists.
   (i) $f(x) = x - 3$,
   (ii) $f(x) = 2x$,
   (iii) $f(x) = \begin{cases} x \text{ if } x \text{ is even} \\ x + 1 \text{ if } x \text{ is odd} \end{cases}$
   for all $x \in \mathbb{Z}$.

7. Let $A = \{1, 2, 3\}$. List all one-one functions from A onto A.

8. Let $A = \{1, 2, \ldots, n\}$. Show that the number of one-one functions of A onto A is $n!$

9. Let $f : A \to B$ be a function. Define a relation R on A by for all $a, b \in A$, aRb if and only if $f(a) = f(b)$. Show that R is an equivalence relation.

10. Given $f : X \to Y$ and $A, B \subseteq X$, prove that
    (i) $f(A \cup B) = f(A) \cup f(B)$,
    (ii) $f(A \cap B) \subseteq f(A) \cap f(B)$,
    (iii) $f(A \backslash B) \subseteq f(A) \backslash f(B)$ if f is one-one.

11. Given $f : X \to Y$. Let $S \subseteq Y$. Define $f^{-1}(S) = \{x \in X \mid f(x) \in S\}$. Let A, $B \subseteq Y$. Prove that
    (i) $f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$,
    (ii) $f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B)$,
    (iii) $f^{-1}(A \backslash B) = f^{-1}(A) \backslash f^{-1}(B)$.

12. Let $f : A \to B$. Let f^* be the inverse relation, i.e.,
$$f^* = \{(y, x) \in B \times A \mid f(x) = y\}.$$
    (i) Show by an example that f^* need not be a function.
    (ii) Show that f^* is a function from $\mathcal{I}(f)$ into A if and only if f is one-one.
    (iii) Show that f^* is a function from B into A if and only if f is one-one and onto B.
    (iv) Show that if f^* is a function from B into A, then $f^{-1} = f^*$.

13. Show that $\mathbb{Z} \sim \mathbb{E}$, where $\mathbb{E}$ is the set of all even integers.

14. Let $A = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$ and $B = \{x \in \mathbb{R} \mid 5 \leq x \leq 8\}$. Show that $f : A \to B$ defined by $f(x) = 5 + (8 - 5)x$ is a one-one function from A onto B.

15. (i) Show that $\mathbb{Z}$ and $3\mathbb{Z}$ are equipollent.
    (ii) Show that $5\mathbb{Z}$ and $7\mathbb{Z}$ are equipollent.

16. Let $S = \{x \in \mathbb{R} \mid 0 < x < 1\}$. Show that $\mathbb{R}^+ \sim S$.

17. (**Schröder-Bernstein**) Let A and B be sets. If $A \sim Y$ for some subset Y of B and $B \sim X$ for some subset X of A, prove that $A \sim B$.

18. Find a one-one mapping from $\mathbb{R}$ onto $\mathbb{R}^+$.

19. Is $\mathbb{Z} \sim \mathbb{Q}$?

20. Let $A = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$ and $B = \{x \in \mathbb{R} \mid 0 < x < 1\}$. Is it true that $A \sim B$?

21. For each of the following statements, write the proof if the statement is true, otherwise give a counterexample.
    (i) A function $f : A \to B$ is one-one if and only if $g \circ f = h \circ f$ for all functions $g, h : B \to A$.
    (ii) A function $f : A \to B$ is one-one if and only if for all subsets C of A, $f(A \backslash C) \supseteq B \backslash f(C)$.

1.5 Binary Operations

The concept of a binary operation is very important in abstract algebra. Throughout the text we will be concerned with sets together with one or more binary operations. In this section, we define binary operations and examine their basic properties.

Definition 1.5.1 *Let S be a nonempty set. A **binary operation** on S is a function from $S \times S$ into S.*

For any ordered pair (x, y) of elements $x, y \in S$, a binary operation assigns a third member of S. For example, $+$ is a binary operation on $\mathbb{Z}$ which assigns 3 to the pair $(2, 1)$.

If $*$ is a binary operation on S, we write $x * y$ for $*(x, y)$, where $x, y \in S$. Since the image of $*$ is a subset of S, we say S is **closed under** $*$.

$\mathbb{Z}$ is closed under $+$ since if we add two integers we obtain an integer. Since $2, 5 \in \mathbb{N}$ and $2 - 5 = -3 \notin \mathbb{N}$, we see that $-$ (subtraction) is not a binary operation of $\mathbb{N}$ and we say that $\mathbb{N}$ is not closed under $-$.

Definition 1.5.2 *A **mathematical system** is an ordered $(n+1)$-tuple $(S, *_1, \ldots, *_n)$, where S is a nonempty set and $*_i$ is a binary operation on S, $i = 1, 2, \ldots, n$. S is called the underlying set of the system.*

Definition 1.5.3 *Let $(S, *)$ be a mathematical system. Then*
*(i) $*$ is called **associative** if for all $x, y, z \in S, x * (y * z) = (x * y) * z$.*
*(ii) $*$ is called **commutative** if for all $x, y \in S, x * y = y * x$.*

Example 1.5.4 *Consider the mathematical system $(\mathbb{Z}, +)$. Since addition of integers is both associative and commutative, $+$ is both associative and commutative.*

Example 1.5.5 *Let A be a nonempty set. Let S be the set of all functions on A, i.e.,*

$$S = \{f \mid f : A \to A\}.$$

Since composition of functions is a function (Theorem 1.4.11*), $(S, \circ)$ is a mathematical system. By Theorem* 1.4.13*, $\circ$ is associative.*

Example 1.5.6 *Let $M_2(\mathbb{R})$ be the set of all 2×2 matrices over $\mathbb{R}$, i.e.,*

$$M_2(\mathbb{R}) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a, b, c, d \in \mathbb{R} \right\}.$$

Let $+$ denote the usual addition of matrices and $\cdot$ denote the usual multiplication of matrices. Since addition (multiplication) of 2×2 matrices over $\mathbb{R}$ is a 2×2 matrix over $\mathbb{R}$, it follows that $+$ ($\cdot$) is a binary operation on $M_2(\mathbb{R})$. Hence, $(M_2(\mathbb{R}), +, \cdot)$ is a mathematical system. Note that $+$ is both associative and commutative and $\cdot$ is associative, but not commutative.

The following is an example of a mathematical system for which the binary operation is neither associative nor commutative.

Example 1.5.7 *Consider the mathematical system $(\mathbb{Z}, -)$, where $-$ denotes the binary operation of subtraction on $\mathbb{Z}$. Then $3-(2-1) = 2 \neq 0 = (3-2)-1$ and so $-$ is not associative. Also, since $3-2 \neq 2-3$, $-$ is not commutative.*

A convenient way to define a binary operation on a finite set S is by means of an operation or multiplication table. For example, let $S = \{a, b, c\}$. Define $*$ on S by the following operation table.

$*$	a	b	c
a	c	b	a
b	a	a	a
c	b	b	b

To determine the element of S assigned to $a * b$, we look at the intersection of the row labeled by a and the column headed by b. We see that $a * b = b$. Note that $b * a = a$.

Definition 1.5.8 *Let $(S, *)$ be a mathematical system. An element $e \in S$ is called an **identity** of $(S, *)$ if for all $x \in S$,*

$$e * x = x = x * e.$$

Example 1.5.9 *Let $S = \{e, a, b\}$. Define $*$ on S by the following multiplication table*

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

We note that $e * a = a = a * e$, $e * b = b = b * e$ and $e * e = e = e * e$. Thus, e is an identity of $(S, *)$.

Example 1.5.10 *(i) In Example* 1.5.5*, i_A is an identity element of $(S, \circ)$.*

(ii) In Example 1.5.6*, $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ is an identity element for the mathematical system $(M_2(\mathbb{R}), +)$ and $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is an identity element for the mathematical system $(M_2(\mathbb{R}), \cdot)$.*

Theorem 1.5.11 *An identity element (if it exists) of a mathematical system $(S, *)$ is unique.*

Proof. Let e, f be identities of $(S, *)$. Since e is identity, $e * a = a$ for all $a \in S$. Substituting f for a, we get

$$e * f = f. \tag{1.4}$$

Now f is identity and so $a * f = a$ for all $a \in S$. Substituting e for a we get

$$e * f = e. \tag{1.5}$$

From Eqs. (1.4) and (1.5), we get $e = f$. Hence, an identity element (if it exists) is unique. ■

Worked-Out Exercises

Exercise 1 Which of the following are associative binary operations?

(i) $(\mathbb{Z}, *)$, where $x * y = (x + y) - (x \cdot y)$ for all $x, y \in \mathbb{Z}$.

(ii) $(\mathbb{R}, *)$, where $x * y = \max(x, y)$ for all $x, y \in \mathbb{R}$.

(iii) $(\mathbb{R}, *)$, where $x * y = |x + y|$ for all $x, y \in \mathbb{R}$.

Solution (i) $(x * y) * z = ((x + y) - (x \cdot y)) * z = (x + y) - (x \cdot y) + z - ((x + y) - (x \cdot y)) \cdot z = x + y + z - x \cdot y - x \cdot z - y \cdot z + x \cdot y \cdot z$. Similarly, $x * (y * z) = x + y + z - x \cdot y - x \cdot z - y \cdot z + x \cdot y \cdot z$. Thus, $(x * y) * z = x * (y * z)$. Hence, $*$ is associative.

(ii) $(x*y)*z = \max(x, y)*z = \max(\max(x, y), z) = \max(x, y, z) = \max(x, \max(y, z)) = x*\max(y, z) = x * (y * z)$. Thus, $*$ is associative.

(iii) $(2 * (-3)) * 6 = |2 + (-3)| * 6 = 1 * 6 = |1 + 6| = 7$ and $2 * ((-3) * 6) = 2 * (|(-3) + 6|) = 2 * 3 = |2 + 3| = 5$. Hence, $(2 * (-3)) * 6 \neq 2 * ((-3) * 6)$ and so $*$ is not associative.

Exercises

1. Which of the following are associative binary operations?

   (i) $(\mathbb{N}, *)$, where $x * y = x^y$ for all $x, y \in \mathbb{N}$.

   (ii) $(\mathbb{Z}, *)$, where $x * y = x + y + 1$ for all $x, y \in \mathbb{Z}$.

   (iii) $(\mathbb{N}, *)$, where $x * y = \gcd(x, y)$ for all $x, y \in \mathbb{N}$.

   (iv) $(\mathbb{N}, *)$, where $x * y = \text{lcm}(x, y)$ for all $x, y \in \mathbb{N}$.

   (v) $(\mathbb{R}, *)$, where $x * y = \min(x, y)$ for all $x, y \in \mathbb{R}$.

   (vi) $(\mathbb{R}, *)$, where $x * y = |x| + |y|$ for all $x, y \in \mathbb{R}$.

2. In Exercise 1, which of the operations are commutative?

3. In Exercise 1, which mathematical systems have an identity?

Chapter 2

Introduction to Groups

There are four major sources from which group theory evolved, namely, classical algebra, number theory, geometry, and analysis. Classical algebra originated in 1770 with J. L. Lagrange's work on polynomial equations. His work appeared in a memoir entitled, "Réflexions sur la résolution algébrique des équations." C. F. Gauss is considered the originator of number theory with his work, "*Disquistiones Arithmeticae*," which was published in 1801. F. Klein's lecture in 1872, "A Comparative Review of Recent Researches in Geometry," dealt with the classification of geometry as the study of invariants under groups of transformations. The impact of his lecture was so strong as to allow Klein to be considered as the originator of this source of group theory. The originators of the analysis source are S. Lie (1874) and H. Poincaré and F. Klein (1876).

2.1 Elementary Properties of Groups

In this chapter, and in fact in the remainder of the text, we will be concerned with mathematical systems. These systems are composed of a nonempty set together with binary operations defined on this set so that certain properties hold. From these properties, results concerning these systems are derived. This axiomatic approach to abstract algebra unifies diverse examples and also strips away nonessential ideas.

Although noted for his geometry, Euclid inspired the use of the axiomatic method, which has proved so indispensable in mathematics. His axiomatic approach also affected philosophy, where in the 17th century Baruch Spinoza laid down (in The *Ethics*) an axiomatic system from which he was able to prove the existence of God. His proof, of course, depended on his axioms. His proof lost its conviction with the emergence of noneuclidean geometries whose axioms were as logical and practical as Euclid's.

We will be primarily concerned with mathematical systems called groups in this chapter. The theory of groups is one of the oldest branches of abstract algebra. The first effective use of groups was in the early nineteenth century by A. Cauchy and E. Galois. They used groups to describe the effect of permutations of roots of a polynomial equation. Their use of groups was not based on an axiomatic approach. In 1854, A. Cayley gave the first postulates for a group. However, his definition was lost sight of. Kronecker again set down the axioms for an Abelian group in 1870. H. Weber gave the definition for finite groups (in 1882) and the definition for infinite groups in 1883.

As previously mentioned, the notion of a group arose from the study of one-one functions on the set of roots of a polynomial equation. We have seen that the set S of all one-one functions from a set X onto itself satisfies the following properties:

(i) Composition of functions, $\circ$, is a binary operation on S.

(ii) For all $f, g, h \in S$, $f \circ (g \circ h) = (f \circ g) \circ h$.

(iii) There exists $i \in S$ such that $f \circ i = f = i \circ f$ for all $f \in S$.

(iv) For all $f \in S$ there exists an element $f^{-1} \in S$ such that $f \circ f^{-1} = i = f^{-1} \circ f$.

These properties lead us to the definition of an abstract group.

Definition 2.1.1 *A **group** is an ordered pair $(G, *)$, where G is a nonempty set and $*$ is a binary operation on G such that the following properties hold:*

*(G1) For all $a, b, c \in G$, $a * (b * c) = (a * b) * c$ (**associative law**).*

*(G2) There exists $e \in G$ such that for all $a \in G$, $a * e = a = e * a$ (**existence of an identity**).*

*(G3) For all $a \in G$, there exists $b \in G$ such that $a * b = e = b * a$ (**existence of an inverse**).*

Thus, a group is a mathematical system $(G, *)$ satisfying axioms G1 to G3.

Example 2.1.2 *Consider $\mathbb{Z}$, the set of integers, together with the binary operation $+$, where $+$ is the usual addition. We know that $+$ is associative. Now $0 \in \mathbb{Z}$ and for all $a \in \mathbb{Z}$,*

$$a + 0 = a = 0 + a.$$

So 0 is an identity. Also, for all $a \in \mathbb{Z}$, $-a \in \mathbb{Z}$ and

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

That is, $-a$ is an inverse of a. It now follows that $(\mathbb{Z}, +)$ satisfies axioms G1 to G3, so $(\mathbb{Z}, +)$ is a group.

As in Example 2.1.2, we can show that $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, $(\mathbb{C}, +)$ are also groups, where $+$ is the usual addition.

Example 2.1.3 *Consider $\mathbb{Q}\backslash\{0\}$, the set of nonzero rational number, together with the binary operation $\cdot$, where $\cdot$ is the usual multiplication. We know that $\cdot$ is associative. Now $1 \in \mathbb{Q}$ and for all $a \in \mathbb{Q}$,*

$$a \cdot 1 = a = 1 \cdot a.$$

So 1 is an identity. Also, for all $a \in \mathbb{Q}\backslash\{0\}$, $\frac{1}{a} \in \mathbb{Q}\backslash\{0\}$ and

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

This implies that $\frac{1}{a}$ is an inverse of a. It now follows that $(\mathbb{Q}\backslash\{0\}, \cdot)$ satisfying axioms G1 to G3, so $(\mathbb{Q}\backslash\{0\}, \cdot)$ is a group.

As in Example 2.1.3, we can show that $(\mathbb{R}\backslash\{0\}, \cdot)$, $(\mathbb{C}\backslash\{0\}, \cdot)$ are also groups, where $\cdot$ is the usual multiplication. Note that for each of the groups $(\mathbb{R}\backslash\{0\}, \cdot)$, $(\mathbb{C}\backslash\{0\}, \cdot)$ the identity element is 1.

Example 2.1.2 shows that 0 is an identity of $(\mathbb{Z}, +)$ and for each element $a \in \mathbb{Z}$, $-a$ is an inverse of a. Similarly, Example 2.1.3 shows that 1 is an identity of $(\mathbb{Q}\backslash\{0\}, \cdot)$ and for each element $a \in \mathbb{Q}\backslash\{0\}$, $\frac{1}{a}$ is an inverse of a. The next theorem shows that in a group there is only one identity element, i.e., identity element is unique. Similarly, in a group, every element has only one inverse, i.e., the inverse of an element is unique.

Theorem 2.1.4 *Let $(G, *)$ be a group.*

(i) There exists a unique element $e \in G$ such that $e * a = a = a * e$ for all $a \in G$.

(ii) For all $a \in G$, there exists a unique $b \in G$ such that $a * b = e = b * a$.

Proof. (i) Now $(G, *)$ is group. Therefore, by G2, there exists $e \in G$ such that $e * a = a = a * e$ for all $a \in G$. Because $(G, *)$ is a mathematical system, e is unique by Theorem 1.5.11.

(ii) Let $a \in G$. By G3, there exists $b \in G$ such that $a * b = e = b * a$. Suppose there exists $c \in G$ such that $a * c = e = c * a$. We show that $b = c$. Now

$$\begin{aligned} b &= b * e \\ &= b * (a * c) && \text{(substituting } e = a * c) \\ &= (b * a) * c && \text{(using the associativity of } *) \\ &= e * c && \text{(because } b * a = e) \\ &= c. \end{aligned}$$

Thus, b is unique. ■

The unique element $e \in G$ that satisfies G2 is called the **identity** element of the group $(G, *)$. Let $a \in G$. Then the unique element $b \in G$ that satisfies G3 is called the **inverse** of a and is denoted by a^{-1}.

Remark 2.1.5 *By Theorem 2.1.4, it follows that for each of the groups* $(\mathbb{Z},+)$, $(\mathbb{Q},+)$, $(\mathbb{R},+)$, *and* $(\mathbb{C},+)$, *the identity element is* 0. *Similarly, for each of the groups* $(\mathbb{Q}\backslash\{0\},\cdot)$, $(\mathbb{R}\backslash\{0\},\cdot)$, *and* $(\mathbb{C}\backslash\{0\},\cdot)$, *the identity element is* 1.

Before giving additional examples of groups, let us make the following definition.

Definition 2.1.6 *Let* $(G,*)$ *be a group. If for all* $a,b\in G$

$$a*b=b*a,$$

then $(G,*)$ *is called a* ***commutative*** *or* ***Abelian*** *group. A group* $(G,*)$ *is called* ***noncommutative*** *if it is not commutative.*

Example 2.1.7 *Consider the group* $(\mathbb{Z},+)$ *of Example 2.1.2. Because* $a+b=b+a$ *for all* $a,b\in\mathbb{Z}$, *it follows that* $+$ *is commutative. Hence,* $(\mathbb{Z},+)$ *is a commutative group.*

Similarly, $(\mathbb{Q},+)$, $(\mathbb{R},+)$, $(\mathbb{C},+)$, $(\mathbb{Q}\backslash\{0\},\cdot)$, $(\mathbb{R}\backslash\{0\},\cdot)$, $(\mathbb{C}\backslash\{0\},\cdot)$ *are also commutative groups, where* $+$ *is the usual addition and* $\cdot$ *is the usual multiplication.*

Next we consider additional examples of (commutative) groups.

Example 2.1.8 *Consider* $\mathbb{Z}$, *the set of integers. Let* a *be a fixed integer. Let*

$$G=\{na\mid n\in\mathbb{Z}\}.$$

That is, G *consists of all multiples of* a. *Note that* $G\subseteq\mathbb{Z}$.

Now $0=0a\in G$. *So it follows that* G *is nonempty. Because* $+$ *is commutative and associative on* $\mathbb{Z}$ *and* G *is a subset of* $\mathbb{Z}$, *it follows that* $+$ *is commutative and associative on* G. *Moreover, note that* 0 *is the identity element of* G. *Also for each* $na\in G$, $-(na)=(-n)a\in G$ *and*

$$na+(-(na))=0=(-(na))+na.$$

We can now conclude that $(G,+)$ *is a commutative group.*

Let n be a fixed positive integer, Chapter 1 extensively describes the set $\mathbb{Z}_n$ and the binary relation $\equiv_n$on $\mathbb{Z}_n$. The next example shows that $\mathbb{Z}_n$ together with the binary relation $+_n$, as defined in that example, is a commutative group. The next two examples are, in fact, due to Gauss's, whose work yielded many new directions of research in Abelian groups.

Example 2.1.9 *Let* n *be a fixed positive integer. Consider* $\mathbb{Z}_n$ *(as defined in Examples* 1.3.11*). Let* $+_n$ *be defined on* $\mathbb{Z}_n$ *by*

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

Recall that $[a]=\{x\in\mathbb{Z}\mid x\equiv_n a\}$. *We show that* $(\mathbb{Z}_n,+_n)$ *is a commutative group.*

First, we show that $+_n$ *is a binary operation on* $\mathbb{Z}_n$. *Let* $[a],[b]\in\mathbb{Z}_n$. *Then* $[a]+_n[b]=[a+b]\in\mathbb{Z}_n$. *Next let,* $[a],[b],[c],[d]\in\mathbb{Z}_n$. *Suppose* $[a]=[c]$ *and* $[b]=[d]$. *Then* $a\equiv c(\mathrm{mod}\ n)$ *and* $b\equiv d(\mathrm{mod}\, n)$. *Thus, there exist* $s,t\in\mathbb{Z}$ *such that*

$$a-c=ns \text{ and } b-d=nt.$$

This implies that

$$a+b-(c+d)=(a-c)+(b-d)=ns+nt=n(s+t).$$

Thus, $a+b\equiv(c+d)(\mathrm{mod}\ n)$, *so* $[a+b]=[c+d]$. *This implies that* $[a]+_n[b]=[c]+_n[d]$. *Consequently,* $+_n$ *is well-defined. It now follows that* $+_n$ *is a binary operation on* $\mathbb{Z}_n$.

For all $[a],[b],[c]\in\mathbb{Z}_n$,

$$\begin{array}{rcll}
([a]+_n[b])+_n[c] & = & [a+b]+_n[c] & \textit{by the definition of } +_n\\
& = & [(a+b)+c] & \textit{by the definition of } +_n\\
& = & [a+(b+c)] & \textit{becasue } + \textit{ is associative on } \mathbb{Z}\\
& = & [a]+_n[b+c] & \textit{by the definition of } +_n\\
& = & [a]+_n([b]+_n[c]) & \textit{by the definition of } +_n .
\end{array}$$

This implies that, $+_n$ is associative.

Now $[0] \in \mathbb{Z}_n$ and for all $[a] \in \mathbb{Z}_n$,

$$[a] +_n [0] = [a + 0] = [a] = [0 + a] = [0] +_n [a].$$

This shows that $[0]$ is the identity element. Also, for all $[a] \in \mathbb{Z}_n$, $[-a] \in \mathbb{Z}_n$ we have

$$[a] +_n [-a] = [a - a] = [0] = [-a + a] = [-a] +_n [a].$$

Thus, $[-a]$ is the inverse of $[a]$. Finally, for all $[a], [b] \in \mathbb{Z}_n$

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

so $+_n$ is commutative. Hence, $(\mathbb{Z}_n, +_n)$ is a commutative group.

Example 2.1.10 *Let n be a fixed positive integer. Consider $\mathbb{Z}_n$ (as defined in Examples 1.3.11). Let $\cdot_n$ be define on $\mathbb{Z}_n$ by for all $[a]$, $[b] \in \mathbb{Z}_n$*

$$[a] \cdot_n [b] = [ab].$$

First we show that $\cdot_n$ is a binary operation on $\mathbb{Z}_n$. Let $[a]$, $[b] \in \mathbb{Z}_n$. Then $[a] \cdot_n [b] = [ab] \in \mathbb{Z}_n$. Next let, $[a], [b], [c], [d] \in \mathbb{Z}_n$. Suppose $[a] = [c]$ and $[b] = [d]$. Then $a \equiv c (\mathrm{mod}\ n)$ and $b \equiv d (\mathrm{mod}\, n)$. Thus, there exist $s, t \in \mathbb{Z}$ such that

$$a - c = ns \text{ and } b - d = nt.$$

This implies that

$$\begin{aligned} ab - cd &= ab - bc + bc - cd = b(a - c) + c(b - d) \\ &= bns + cnt = n(bs + ct). \end{aligned}$$

Thus, $ab \equiv cd (\mathrm{mod}\ n)$, so $[ab] = [cd]$. This implies that $[a] \cdot_n [b] = [c] \cdot_n [d]$. Consequently, $\cdot_n$ is well-defined. It now follows that $\cdot_n$ is a binary operation on $\mathbb{Z}_n$. Thus, $(\mathbb{Z}_n, \cdot_n)$ is a mathematical system.

Moreover, as in Example 2.1.9, we can show that $\cdot_n$ is associative.

Now $[1] \in \mathbb{Z}_n$ and for all $[a] \in \mathbb{Z}_n$,

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

This implies that $[1]$ is the identity element.

Let $[a] \in \mathbb{Z}_n$ and $[a] \neq [0]$. We leave it as an exercise to show that $[a]$ has an inverse if and only if $\gcd(a, n) = 1$. Thus, we see that in general, not every element of $\mathbb{Z}_n \backslash \{[0]\}$ has an inverse.

For example if $n = 6$, then the only elements of $\mathbb{Z}_6$ that have inverses are $[1]$, $[3]$, and $[5]$. Hence, in general $(\mathbb{Z}_n \backslash \{[0]\}, \cdot_n)$ is not a group.

Let U_n be the set of all elements of $\mathbb{Z}_n \backslash \{[0]\}$ that have an inverse in $(\mathbb{Z}_n \backslash \{[0]\}, \cdot_n)$, i.e.,

$$U_n = \{[a] \in \mathbb{Z}_n \backslash \{[0]\} \mid \gcd(a, n) = 1\}.$$

We ask the reader to verify in Exercise 10 (page 55) that $(U_n, \cdot_n)$ is a group.

Note that for $n = 8$, $U_8 = \{[1], [3], [5], [7]\}$ and for $n = 7$,

$$U_7 = \{[1],\ [2],\ [3],\ [4],\ [5],\ [6]\} = \mathbb{Z}_7 \backslash \{[0]\}.$$

Example 2.1.11 *Let*

$$\mathbb{Q}[\sqrt{2}] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}.$$

Note that $\mathbb{Q}[\sqrt{2}] \subseteq \mathbb{R}$. Now $0 = 0 + 0\sqrt{2} \in \mathbb{Q}[\sqrt{2}]$. This shows that $\mathbb{Q}[\sqrt{2}] \neq \emptyset$. Define $+$ on $\mathbb{Q}[\sqrt{2}]$ as follows: for all $a + b\sqrt{2}, c + d\sqrt{2} \in \mathbb{Q}[\sqrt{2}]$,

$$(a + b\sqrt{2}) + (c + d\sqrt{2}) = (a + b) + (c + d)\sqrt{2}).$$

It is easy to see that + is a binary operation on $\mathbb{Q}[\sqrt{2}]$. Note that + is the usual addition. Because $\mathbb{Q}[\sqrt{2}] \subseteq \mathbb{R}$, it follows that + is associative and commutative on $\mathbb{Q}[\sqrt{2}]$. Next, for all $a + b\sqrt{2} \in \mathbb{Q}[\sqrt{2}]$,

$$(a + b\sqrt{2}) + (0 + 0\sqrt{2}) = a + b\sqrt{2} = (0 + 0\sqrt{2}) + (a + b\sqrt{2}).$$

Thus, $0 = 0+0\sqrt{2}$ is the identity element of $(\mathbb{Q}[\sqrt{2}], +)$. Note that the inverse of $a+b\sqrt{2}$ is $-a+(-b)\sqrt{2}$. Hence, $(\mathbb{Q}[\sqrt{2}], +)$ is a commutative groups.

In a similar manner, we can show that $(\mathbb{Q}[\sqrt{2}]\backslash\{0\}, \cdot)$ is a commutative group, where $\cdot$ is the usual multiplication. Note that the identity of $(\mathbb{Q}[\sqrt{2}]\backslash\{0\}, \cdot)$ is $1 = 1 + 0\sqrt{2}$ and the inverse of $a + b\sqrt{2} \neq 0$ is $\frac{a}{a^2-2b^2} - \frac{b}{a^2-2b^2}\sqrt{2}$.

Example 2.1.12 *Let $\mathcal{P}(X)$ be the power set of a set X. Consider the operation Δ on $\mathcal{P}(X)$. Then for all A, $B \in \mathcal{P}(X)$,*

$$A\Delta B = (A\backslash B) \cup (B\backslash A).$$

It can be verified that Δ is a binary operation on $\mathcal{P}(X)$ and Δ is associative.

Now $\emptyset \in \mathcal{P}(X)$. Let $A \in \mathcal{P}(X)$. Then

$$\begin{array}{rcll} A\Delta\emptyset & = & (A\backslash\emptyset) \cup (\emptyset\backslash A) & \\ & = & A \cup \emptyset & \text{because } A\backslash\emptyset = A \text{ and } \emptyset\backslash A = \emptyset \\ & = & A. & \end{array}$$

Similarly, $\emptyset\Delta A = A$. Thus, $A\Delta\emptyset = A = \emptyset\Delta A$. It now follows that $\emptyset$ is the identity element.

Next,

$$\begin{array}{rcll} A\Delta A & = & (A\backslash A) \cup (A\backslash A) & \\ & = & \emptyset \cup \emptyset & \text{because } A\backslash A = \emptyset \text{ and } A\backslash A = \emptyset \\ & = & \emptyset. & \end{array}$$

This implies that A is the inverse of A, i.e., A is its own inverse. We can now conclude that $(\mathcal{P}(X), \Delta)$ is a group.

We also note that for all $A, B \in \mathcal{P}(X)$

$$\begin{array}{rcl} A\Delta B & = & (A\backslash B) \cup (B\backslash A) \\ & = & (B\backslash A) \cup (A\backslash B) \\ & = & B\Delta A. \end{array}$$

This shows that Δ is commutative on $\mathcal{P}(X)$. Consequently, $(\mathcal{P}(X), \Delta)$ is a commutative group.

Example 2.1.13 *Let X be a set and S_X the set of all one-one functions of X onto X, i.e.,*

$$S_X = \{f : X \to X \mid f \text{ is a one-one function of } X \text{ onto } X\}.$$

Because i_X, the identity function on X, is one-one and onto X, $i_X \in S_X$. Thus, $S_X \neq \emptyset$.

Let $f, g \in S_X$. Then $f \circ g$ is a one-one function of X onto X by Theorem 1.4.11. Hence, $f \circ g \in S_X$. By Theorem 1.4.13, $\circ$ is associative. Also, for all $f \in S_X$, $f^{-1} \in S_X$ and $f \circ f^{-1} = i_X = f^{-1} \circ f$. Consequently, $(S_X, \circ)$ is a group.

Note that, $(S_X, \circ)$ is not necessarily commutative. For example, let $X = \{a, b, c\}$. Let $f, g \in S_X$ be defined by

$$f(a) = b, f(b) = a, f(c) = c,$$

$$g(a) = b, g(b) = c, g(c) = a.$$

Then

$$(f \circ g)(b) = f(g(b)) = f(c) = c$$

and

$$(g \circ f)(b) = g(f(b)) = g(a) = b.$$

This implies that $(f \circ g)(b) \neq (g \circ f)(b)$. Hence, $f \circ g \neq g \circ f$. Thus, $(S_X, \circ)$ is not commutative.

Example 2.1.14 *Let* $GL(2,\mathbb{R}) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a,b,c,d \in \mathbb{R},\ ad-bc \neq 0 \right\}$. *Define a binary operation* $*$ *on* $GL(2,\mathbb{R})$ *by*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} * \begin{bmatrix} u & v \\ w & s \end{bmatrix} = \begin{bmatrix} au+bw & av+bs \\ cu+dw & cv+ds \end{bmatrix}$$

for all $\begin{bmatrix} a & b \\ c & d \end{bmatrix}, \begin{bmatrix} u & v \\ w & s \end{bmatrix} \in GL(2,\mathbb{R})$. *This binary operation is the usual matrix multiplication. Note that* $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \in GL(2,\mathbb{R})$. *Thus,* $GL(2,\mathbb{R}) \neq \emptyset$.

Because matrix multiplication is associative, $*$ *is associative. Next consider the matrix* $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. *Because* $1\cdot 1 - 0\cdot 0 = 1 \neq 0$, *it follows that* $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \in GL(2,\mathbb{R})$. *Moreover, for any* $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in GL(2,\mathbb{R})$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} * \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

This implies that $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ *is the identity element of* $GL(2,\mathbb{R})$.

Let $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in GL(2,\mathbb{R})$. *Then* $ad-bc \neq 0$. *Consider the matrix* $\begin{bmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix}$. *Because*

$$\frac{d}{ad-bc}\cdot\frac{a}{ad-bc} - \frac{-b}{ad-bc}\cdot\frac{-c}{ad-bc} = \frac{1}{ad-bc} \neq 0,$$

we have

$$\begin{bmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix} \in GL(2,\mathbb{R}).$$

Now

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} * \begin{bmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and

$$\begin{bmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix} * \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus, $\begin{bmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix}$ *is the inverse of* $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$. *Hence,* $(GL(2,\mathbb{R}),*)$ *is a group.*

Next note that

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \in GL(2,\mathbb{R})$$

and

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \neq \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Hence, $(GL(2,\mathbb{R}),*)$ *is a noncommutative group.*

Remark 2.1.15 *The group in Example 2.1.14 is known as the* ***general linear group of degree 2****.*

We now prove some elementary properties of a group in the following theorem.

Theorem 2.1.16 *Let* $(G,*)$ *be a group.*
(i) $(a^{-1})^{-1} = a$ *for all* $a \in G$.
(ii) $(a*b)^{-1} = b^{-1}*a^{-1}$ *for all* $a,b \in G$.
*(iii) (****Cancellation Law****) For all* $a,b,c \in G$, *if either* $a*c = b*c$ *or* $c*a = c*b$, *then* $a = b$.
(iv) For all $a,b \in G$, *the equations* $a*x = b$ *and* $y*a = b$ *have unique solutions in* G *for* x *and* y.

Proof. (i) Let $a \in G$. Then $a^{-1} * a = e = a * a^{-1}$. Thus, by the definition of an inverse of an element, a is an inverse of a^{-1}. However, by Theorem 2.1.4, the inverse of an element is unique in a group. Therefore, because $(a^{-1})^{-1}$ denotes the inverse of a^{-1}, it follows that $(a^{-1})^{-1} = a$.

(ii) Let $a, b \in G$. Then

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

Similarly, $(b^{-1} * a^{-1}) * (a * b) = e$. Hence, $b^{-1} * a^{-1}$ is an inverse of $a * b$. Because the inverse of an element is unique in a group and $(a * b)^{-1}$ denotes the inverse of $a * b$, it follows that $(a * b)^{-1} = b^{-1} * a^{-1}$.

(iii) Let $a, b, c \in G$. Suppose $a * c = b * c$. Now

$$\begin{aligned}& (a * c) * c^{-1} = (b * c) * c^{-1} \\ \Rightarrow\ & a * (c * c^{-1}) = b * (c * c^{-1}) && \text{(by the associativity of } *) \\ \Rightarrow\ & a * e = b * e && \text{(becasue } c * c^{-1} = e) \\ \Rightarrow\ & a = b.\end{aligned}$$

Similarly, if $c * a = c * b$, then we can show that $a = b$.

(iv) Let $a, b \in G$. First we consider the equation

$$a * x = b.$$

Now $a^{-1} * b \in G$. Substituting $a^{-1} * b$ for x in the equation $a * x = b$, we obtain

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

This implies that $a^{-1} * b$ is a solution of the equation $a * x = b$.

We now establish the uniqueness of the solution. Suppose c is any solution of $a * x = b$. Then $a * c = b$. Hence,

$$\begin{aligned}c &= e * c \\ &= (a^{-1} * a) * c && \text{(because } a^{-1} * a = e) \\ &= a^{-1} * (a * c) && \text{(because } * \text{ is associative)} \\ &= a^{-1} * b && \text{(because } a * c = b).\end{aligned}$$

This yields the uniqueness of the solution.

Similar arguments hold for the equation $y * a = b$. ■

Corollary 2.1.17 *Let $(G, *)$ be a group and $a \in G$. If $a * a = a$, then $a = e$.*

Proof. Suppose $a = a * a$. This implies that $a * a = a * e$. By the cancellation law, $a = e$. ■

Corollary 2.1.18 *In a multiplication table for a group $(G, *)$, each element appears exactly once in each row and exactly once in each column.*

Proof. Let $b \in G$ be such that b occurs twice in the row labeled by $a \in G$. Then there exists $u, v \in G$ with $u \neq v$ such that $a * u = b$ and $a * v = b$. This implies that the equation $a * x = b$ has (at least) two distinct solutions, which are u and v. This is a contradiction to Theorem 2.1.16(iv) because the equation $a * x = b$ has a unique solution for x.

A similar argument for columns can be used. ■

Let $(G, *)$ be a group and $a, b, c \in G$. Then by the associative law,

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

Hence, we can define

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

Let $a, b, c, d \in G$. Then

$$\begin{aligned}(a * b * c) * d &= (a * (b * c)) * d \\ &= a * ((b * c) * d)) \\ &= a * (b * (c * d)) \\ &= (a * b) * (c * d) \\ &= ((a * b) * c) * d.\end{aligned}$$

Thus, there is more than one way of inserting parentheses in the expression $a * b * c * d$ to produce a "meaningful product" of a, b, c, d (in this order). We now extend this notion to any finite number of elements.

Definition 2.1.19 *Let $(G, *)$ be a group and $a_1, a_2, \ldots, a_n \in G$ be n elements of G (not necessarily distinct). The* ***meaningful product*** *of $a_1, a_2, \ldots, a_n$ (in this order) is defined as follows:*

If $n = 1$, then the meaningful product is a_1.

If $n > 1$, then the meaningful product of $a_1, a_2, \ldots, a_n$ is any product of the form

$$(a_1 * \cdots * a_m) * (a_{m+1} * \cdots * a_n),$$

where $1 \leq m < n$ and $(a_1 * \cdots * a_m)$ and $(a_{m+1} * \cdots * a_n)$ are meaningful products of m and $n - m$ elements, respectively.

Definition 2.1.20 *Let $(G, *)$ be a group and $a_1, a_2, \ldots, a_n \in G$, $n \geq 1$. The* ***standard product*** *of $a_1, a_2, \ldots, a_n$ denoted by $a_1 * a_2 * \cdots * a_n$ is defined recursively as*

$$\begin{aligned}a_1 &= a_1 \\ a_1 * a_2 * \cdots * a_n &= (a_1 * a_2 * \cdots * a_{n-1}) * a_n \text{ if } n > 1.\end{aligned}$$

In the next theorem, we establish the equality between any meaningful product and standard product.

Theorem 2.1.21 *Let $(G, *)$ be a group and $a_1, a_2, \ldots, a_n \in G$, $n \geq 1$. Then all possible meaningful products of $a_1, a_2, \ldots, a_n$ (in this order) are equal to the standard product of $a_1, a_2, \ldots, a_n$ (in this order).*

Proof. We prove the result by induction.

Basis step: If $n = 1$, then a_1 is the only meaningful product of a_1, which is equal to the standard product a_1 of a_1. Thus, the result is true if $n = 1$.

Inductive hypothesis: Suppose that the theorem is true for all integers m such that $1 \leq m < n$.

Inductive step: Let $a_1, a_2, \ldots, a_n \in G$. Let $(a_1 * \cdots * a_t) * (a_{t+1} * \cdots * a_n)$ be a meaningful product of $a_1, a_2, \ldots, a_n$ (in this order). Now $t < n$ and $n - t < n$. If $t = n - 1$, then

$$(a_1 * a_2 * \cdots * a_t) * a_{t+1} = a_1 * a_2 * \cdots * a_t * a_{t+1}.$$

Suppose $t < n - 1$. Then

$$\begin{aligned}& (a_1 * \cdots * a_t) * (a_{t+1} * \cdots * a_n) \\ = \; & (a_1 * \cdots * a_t) * ((a_{t+1} * \cdots * a_{n-1}) * a_n) \\ = \; & ((a_1 * \cdots * a_t) * (a_{t+1} * \cdots * a_{n-1})) * a_n \\ = \; & (a_1 * a_2 * \cdots * a_{n-1}) * a_n \qquad \text{by the inductive hypothesis,} \\ & \qquad\qquad\qquad\qquad\quad (a_1 * \cdots * a_t) * (a_{t+1} * \cdots * a_{n-1}) \\ & \qquad\qquad\qquad\qquad\quad = a_1 * a_2 * \cdots * a_{n-1}. \\ = \; & a_1 * \cdots * a_n\end{aligned}$$

This shows that the result is true for n. The result now follows by induction. ■

We have seen several examples of groups. In order to show that a given set with a given binary operation is a group, we need to verify G1 to G3 of Definition 2.1.1. However, it would be helpful if we had some criteria that could be used to show whether a given set with a binary operation is a group or not instead of verifying all the properties G1–G3 explicitly. Partly for this reason we define what a semigroup is. Following the examples, we develop some results that can be used to test whether a given set with a binary operation is a group or not.

Definition 2.1.22 *A **semigroup** is an ordered pair $(S, *)$, where S is a nonempty set and $*$ is an associative binary operation on S.*

Thus, a semigroup is a mathematical system with one binary operation such that the binary operation is associative.

Remark 2.1.23 *For any group $(G, *)$, the binary operation $*$ is associative. Therefore, every group $(G, *)$ is a semigroup.*

As in the case of a group, next we define a commutative semigroup.

Definition 2.1.24 *A semigroup $(S, *)$ is **commutative** if $*$ is commutative, i.e., $a * b = b * a$ for all $a, b \in S$. A semigroup $(S, *)$ which is not commutative is called **noncommutative.***

Definition 2.1.25 *Let $(S, *)$ be a semigroup.*
*(i) We say that $(S, *)$ is with **identity** if the mathematical system $(S, *)$ has an identity.*
*(ii) An element $a \in S$ is called **idempotent** if $a * a = a$.*

Example 2.1.26 *(i) Consider $\mathbb{N}$, the set of positive integers. We know that the sum of positive integers is again a positive integer. Thus, $+$ is a binary operation on $\mathbb{N}$. We also know that $+$ is associative and commutative. Thus, $(\mathbb{N}, +)$ is a commutative semigroup. In a similar manner, $(\mathbb{N}, \cdot)$ is a commutative semigroup, where $\cdot$ denotes the usual multiplication of integers.*
(ii) Because $(\mathbb{Z}, +)$ is a commutative group, it is a commutative semigroup. Also note that $(\mathbb{Z}, \cdot)$ is a commutative semigroup.

Example 2.1.27 *Let X be a nonempty set and S the set of all functions $f : X \to X$. If $\circ$ denotes the composition of functions, then $(S, \circ)$ is a semigroup with identity. The associativity of $\circ$ follows from Theorem* 1.4.13.
When X has two or more elements, the semigroup $(S, \circ)$ is noncommutative. For example, let $X = \{a, b\}$. Let $g, h \in S$ be defined by

$$g(a) = b,\ g(b) = b,\ h(a) = b,\ h(b) = a.$$

Then

$$(g \circ h)(a) = b \neq a = (h \circ g)(a).$$

Therefore, $g \circ h \neq h \circ g$.
Let $f \in S$ be defined by $f(a) = a$ and $f(b) = a$. Now

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

for all $x \in G$. Hence, $f \circ g = f \circ h$. But $g \neq h$. This shows that the cancellation laws do not hold in S. Thus, $(S, \circ)$ is not a group.

Example 2.1.28 *Let X be a set with two or more elements and S' the set of all functions $f : X \to X$ which are not one-one. Then $(S', \circ)$ is a noncommutative semigroup without identity.*

Example 2.1.29 *Let X be a set and $\mathcal{P}(X)$ the power set of X. We leave it as an exercise that $(\mathcal{P}(X), \cup)$ and $(\mathcal{P}(X), \cap)$ are commutative semigroups with identity. The identity of $(\mathcal{P}(X), \cup)$ is $\emptyset$ and the identity of $(\mathcal{P}(X), \cap)$ is X.*

The following three theorems give necessary and sufficient conditions for a semigroup to be a group.

Theorem 2.1.30 *A semigroup $(S, *)$ is a group if and only if*
(i) there exists $e \in S$ such that $e * a = a$ for all $a \in S$, (i.e., e is a left identity), and
(ii) for all $a \in S$ there exists $b \in S$ such that $b * a = e$, (i.e., every element has a left inverse).

Proof. Suppose $(S, *)$ is a semigroup that satisfies (i) and (ii). Let a be any element of S. Then there exists $b \in S$ such that $b * a = e$ by (ii). For $b \in S$, there exists $c \in S$ such that $c * b = e$ by (ii). Now

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

and

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

Hence, $a * b = e = b * a$. Also,

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

Thus, $a * e = a = e * a$. This shows that e is the identity element of S. Now because $a * b = e = b * a$, we have $b = a^{-1}$. Therefore, $(S, *)$ is a group.

The converse follows from the definition of a group. ■

Remark 2.1.31 *To verify that a specific nonempty set is a group, we can use Theorem 2.1.30 as follows: Show that (1) the operation, say $*$, defined on the set is well-defined; (2) $*$ is associative; (3) the set has a left identity; and (4) every element has a left inverse. For example, see Worked-Out Exercises 1, 2, and 3 at the end of this section.*

Remark 2.1.32 *The analog of Theorem 2.1.30 is given in Exercise 39 at the end of this section.*

Theorem 2.1.33 *A semigroup $(S, *)$ is a group if and only if for all $a, b \in S$ the equations $a * x = b$ and $y * a = b$ have solutions in S for x and y.*

Proof. Suppose the given equations have solutions in S. Let $a \in S$. Consider the equation $y * a = a$. By our assumption, $y * a = a$ has a solution $u \in S$, say. Then $u * a = a$. Let b be any element of S. Consider the equation $a * x = b$. Again by our assumption, $a * x = b$ has a solution in S. Let $c \in S$ be a solution of $a * x = b$. Then $a * c = b$. Now

$$\begin{aligned} u * b &= u * (a * c) && \text{(because } b = a * c\text{)} \\ &= (u * a) * c && \text{(because } * \text{ is asociative)} \\ &= a * c && \text{(because } u * a = a\text{)} \\ &= b. \end{aligned}$$

Because b was an arbitrary element of S, we find that $u * b = b$ for all $b \in S$. Thus, $(S, *)$ satisfies (i) of Theorem 2.1.30. Consider the equation $y * a = u$. Let $d \in S$ be a solution of $y * a = u$. Then $d * a = u$. This shows that $(S, *)$ satisfies (ii) of Theorem 2.1.30. Hence, $(S, *)$ is a group by Theorem 2.1.30.

The converse follows by Theorem 2.1.16(iv). ■

The next theorem gives a necessary and sufficient condition for a finite semigroup to be a group.

Theorem 2.1.34 *A finite semigroup $(S, *)$ is a group if and only if $(S, *)$ satisfies the cancellation laws (i.e., $a * c = b * c$ implies $a = b$ and $c * a = c * b$ implies $a = b$ for all $a, b, c \in S$).*

Proof. Let $(S, *)$ be a finite semigroup satisfying the cancellation laws. Let $a, b \in S$. Consider the equation $a * x = b$. We show that this equation has a solution in S.

Let us write $S = \{a_1, a_2, \ldots, a_n\}$, where the a_i's are all distinct elements of S. Because S is a semigroup, $a * a_i \in S$ for all $i = 1, 2, \ldots, n$. Thus,

$$\{a * a_1, a * a_2, \ldots, a * a_n\} \subseteq S.$$

Suppose $a * a_i = a * a_j$ for some $i \neq j$. Then by the cancellation law we have $a_i = a_j$, which is a contradiction because $a_i \neq a_j$. Hence, all elements in $\{a * a_1, a * a_2, \ldots, a * a_n\}$ are distinct. Thus,

$$S = \{a * a_1, a * a_2, \ldots, a * a_n\}.$$

Let $b \in S$. Then $b = a * a_k$ for some $a_k \in S$. Therefore, the equation $a * x = b$ has a solution in S. Similarly, we can show that the equation $y * a = b$ has a solution in S. Hence, by Theorem 2.1.33, $(S, *)$ is a group.

The converse follows by Theorem 2.1.16(iii). ■

Let $(G, *)$ be a group, $a \in G$, and $n \in \mathbb{Z}$. We now define the **integral power** a^n of a as follows:

$$\begin{aligned} a^0 &= e \\ a^n &= a * a^{n-1} \text{ if } n > 0 \\ a^n &= (a^{-1})^{-n} \text{ if } n < 0. \end{aligned}$$

Note that $a^n = (a^{-n})^{-1}$ if $n < 0$.

Example 2.1.35 *Let $X = \{a, b, c\}$ be a nonempty set and S_X be the set of all one-one functions from X onto X. Then as in Example , $(S_X, \circ)$ is a group, where $\circ$ is the composition of functions. Consider the function $f : X \to X$ such that*

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

Note that f is a one-one function of X onto X, so $f \in S_X$. Now $f^2 = f \circ f$. Let us determine f^2. We have

$$f^2(a) = (f \circ f)(a) = f(f(a)) = f(b) = c.$$

Similarly, $f^2(b) = a$ and $f^2(c) = b$. Thus, $f^2 \in S_X$ is defined by $f^2(a) = c$, $f^2(b) = a$ and $f^2(c) = b$.

It should be pointed out that when we use additive notation for the binary operation $*$, we speak of multiples of an element a of the group $(G, +)$, which are defined as follows:

$$\begin{aligned} 0a &= 0, \text{ where the 0 on the right-hand side denotes the identity of the} \\ &\quad \text{group } (G, +) \text{ and the 0 on the left-hand side denotes the integer 0.} \\ na &= a + (n-1)a \quad \text{if } n > 0. \\ na &= (-n)(-a) \quad \text{if } n < 0. \end{aligned}$$

Note that, by the notation na, we do not mean n and a multiplied together because no multiplicative operation between elements of $\mathbb{Z}$ and G has been defined.

Example 2.1.36 *Consider the group $(\mathbb{Z}_6, +_6)$ and $[3] \in \mathbb{Z}_6$. We have*

$$2[3] = [3] +_6 [3] = [6] = [0].$$

Similarly,

$$3[5] = [5] +_6 2[5] = [5] +_6 ([5] +_6 [5]) = [5] +_6 [10] = [15] = [3].$$

Note that $15 = 2 \cdot 6 + 3$, so $[15] = [3]$.

Remark 2.1.37 *Let n be a fixed positive integer. Consider the group $(\mathbb{Z}_n, +_n)$. Let $[a] \in \mathbb{Z}_n$. For any integer k, by induction, we can show that*

$$k[a] = [ka].$$

In the exercises at the end of this section, we ask the reader to verify certain basic properties of integral powers.

Definition 2.1.38 *A group $(G, *)$ is called a **finite group** if G has only a finite number of elements. The **order,** written $|G|$, of a group $(G, *)$ is the number of elements of G.*

Remark 2.1.39 *Example 2.1.9 shows that for every positive integer n, there is a commutative group of order n.*

The groups in Examples 2.1.9 and 2.1.10 are finite groups.

Definition 2.1.40 *A group with an infinite number of elements is called an* ***infinite group****.*

Klein and Lie's use of groups in geometry influenced the turn from finite groups to infinite groups.

Example 2.1.41 *The groups in Examples 2.1.7, 2.1.8, and 2.1.11 are infinite groups.*

Let G be a finite group and $a \in G$. Now $a^2 = a*a \in G$ and by induction, we can show that $a^m \in G$ for all $m \geq 1$. Thus, $\{a, a^2, \ldots, a^m, \ldots\} \subseteq G$. Because G is finite, all elements of the set $\{a, a^2, \ldots, a^m, \ldots\}$ cannot be distinct. Hence, $a^k = a^l$ for some positive integers k, l, $k > l$. This implies that $a^{k-l} = e$. Let us write $n = k - l$. Therefore, $a^n = e$ for some positive integer n. Also, if G is an infinite group and $a \in G$, then it may still be possible that $a^n = e$ for some positive integer n. This leads us to the following definition.

Definition 2.1.42 *Let $(G, *)$ be a group and $a \in G$. If there exists a positive integer n such that $a^n = e$, then the smallest such positive integer is called the* ***order*** *of a. If no such positive integer n exists, then we say that a is of* ***infinite order****.*

Notation 2.1.43 *We denote the order of an element a of a group $(G, *)$ by $\circ(a)$.*

The concept of the order of an element is very important in group theory. We shall see in later chapters how effectively information about the order of an element of a group reveals the nature of the group and in several instances leads us to determine the structure of the group itself.

Example 2.1.44 *Consider the group $(\mathbb{Z}_6, +_6)$. $\mathbb{Z}_6$ has order 6. Consider the element $[1]$. Now*

$$\begin{array}{ll} 1[1] = [1] \neq [0], & 2[1] = [2] \neq [0], \\ 3[1] = [3] \neq [0] & 4[1] = [4] \neq [0] \\ 4[1] = [4] \neq [0] & 5[1] = [5] \neq [0] \\ 6[1] = [6] = [0]. & \end{array}$$

This implies that 6 is the smallest positive integer such that $6[1] = [0]$. Hence, $\circ([1]) = 6$. For $[2]$ we have

$$1[2] = [2] \neq [0], \qquad 2[2] = [4] \neq [0], \qquad 3[2] = [6] = [0].$$

That is, 3 is the smallest positive integer such that $3[2] = [0]$. Hence, $\circ([2]) = 3$.

In a similar manner, we can show that $\circ([0]) = 1$, $\circ([3]) = 2$, $\circ([4]) = 3$, and $\circ([5]) = 6$.

Example 2.1.45 *Consider the group $G(2, \mathbb{R})$ of Example 2.1.14. Also consider the elements $\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. Note that both these elements are in $G(2, \mathbb{R})$. Now*

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

This implies that the order of $\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ is 2. Next, we consider $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. Here

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^2 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^3 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^2 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}.$$

By induction, we can show that

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^n = \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix} \text{ for all positive inetegers } n.$$

This implies that the order of $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ is infinite.

Let G be a group and $a \in G$. If $\circ(a)$ is infinite, then by the definition of the order of an element it follows that $\circ(a^k)$ is also infinite for all $k \geq 1$, i.e., the order of every positive power of a is also infinite. If $\circ(a)$ is finite, then the next theorem tells us how to compute the order of various powers of a.

Theorem 2.1.46 *Let $(G, *)$ be a group and a be an element of G such that $\circ(a) = n$.*

(i) If $a^m = e$ for some positive integer m, then n divides m.

(ii) For every positive integer t,

$$\circ(a^t) = \frac{n}{\gcd(t, n)}.$$

Proof. (i) By the division algorithm, there exist $q, r \in \mathbb{Z}$ such that $m = nq + r$, where $0 \leq r < n$. Now

$$a^r = a^{m-nq} = a^m * a^{-nq} = a^m * (a^n)^{-q} = e * (e)^{-q} = e.$$

Now $\circ(a) = n$. Therefore, n is the smallest positive integer such that $a^n = e$. However, $a^r = e$ and $0 \leq r < n$. Thus, we must have that $r = 0$. This implies that $m = nq$. Hence, n divides m.

(ii) Let $\circ(a^t) = k$. Then $a^{kt} = e$. By (i), n divides kt. Thus, there exists $r \in \mathbb{Z}$ such that $kt = nr$.

Let $\gcd(t, n) = d$. Then there exist integers u and v such that

$$t = du, \ n = dv, \text{ and } \gcd(u, v) = 1$$

by Exercise 9 (page 16).

Now $kt = nr$ implies that $kdu = dvr$. Thus, $ku = rv$, i.e., v divides ku. Now $\gcd(u, v) = 1$ and v divides ku. So v divides k. Because $\frac{n}{d} = v$, we have $\frac{n}{d}$ divides k.

Now

$$(a^t)^{\frac{n}{d}} = a^{\frac{nt}{d}} = a^{\frac{ndu}{d}} = a^{nu} = (a^n)^u = e^u = e.$$

We therefore have $\circ(a^t) = k$ and $(a^t)^{\frac{n}{d}} = e$. Therefore, as in (i), k divides $\frac{n}{d}$. Because k and $\frac{n}{d}$ are positive integers such that k divides $\frac{n}{d}$ and $\frac{n}{d}$ divides k, we must have $k = \frac{n}{d}$. Hence,

$$\circ(a^t) = k = \frac{n}{d} = \frac{n}{\gcd(t, n)}.$$

■

Definition 2.1.47 *A group $(G, *)$ is called a **torsion group** if every element of G is of finite order. If every nonidentity element of G is of infinite order, then G is called a **torsion-free group.***

Exercise 2.1.48 *(i) The group of Example 2.1.44 is a torsion group.*

(ii) The groups $(\mathbb{R}, +)$, $(\mathbb{R}^+, \cdot)$, $(\mathbb{Q}^+, \cdot)$ are torsion-free groups.

(iii) Consider the group $(\mathbb{R}\backslash\{0\}, \cdot)$. Now $(-1)^2 = 1$, so $\circ(-1) = 2$. However, if $x \in \mathbb{R}\backslash\{0\}$ such that $x \neq 1$ and $x \neq -1$, then x is of infinite order. It now follows that the groups $(\mathbb{R}\backslash\{0\}, \cdot)$ has elements of finite as well as infinite orders. Hence, $(\mathbb{R}\backslash\{0\}, \cdot)$ is neither a torsion group nor a torsion-free group.

We close this chapter with the following example. The ideas set forth in this example are due to Klein.

Example 2.1.49 *Imagine a square having its sides parallel to the axes of a coordinate system and its center at the origin.*

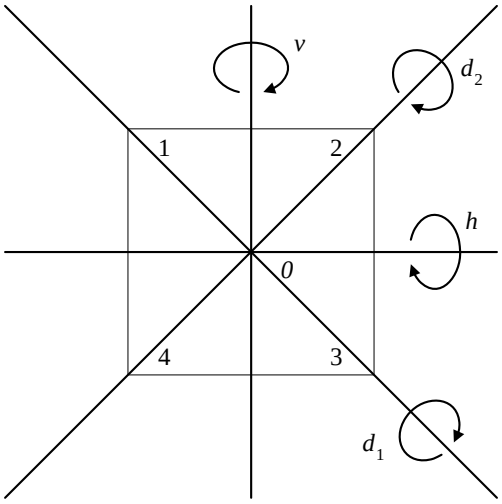


Figure 2-1 Rigid motions of a square

We label the vertices as in the figure and we allow the following rigid motions of the square: clockwise rotations of the square about the center and through angles of 90°, 180°, 270°, 360°, *say,* r_{90}, r_{180}, r_{270}, r_{360}, *respectively; reflections* h *and* v *about the horizontal and vertical axes; reflections* d_1, d_2 *about the diagonals. The diagrams in Figure 2-2 should prove helpful.*

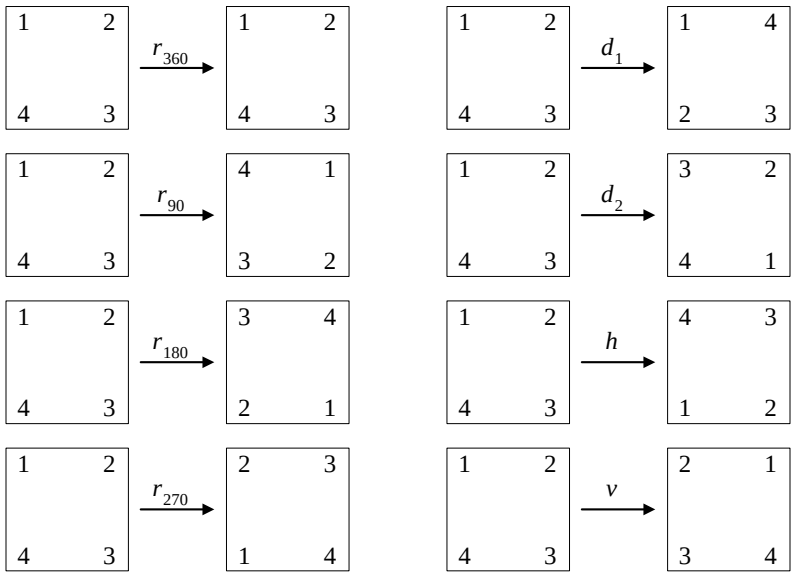


Figure 2-2 Rigid motions of a square

A multiplication $*$ *on two rigid motions can be defined by performing two such motions in succession. For example,* $r_{90} * h$ *is determined by first performing motion* h *and then the motion* r_{90}, *see Figure 2-3.*

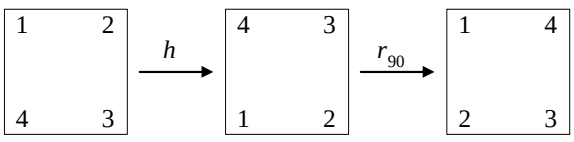


Figure 2-3 Motion h *followed by the motion* r_{90}; $r_{90} * h$

In Figure 2-3, we see that $r_{90} * h = d_1$. *The complete multiplication table for the operation* $*$ *follows.*

$*$	r_{360}	r_{90}	r_{180}	r_{270}	h	v	d_1	d_2
r_{360}	r_{360}	r_{90}	r_{180}	r_{270}	h	v	d_1	d_2
r_{90}	r_{90}	r_{180}	r_{270}	r_{360}	d_1	d_2	v	h
r_{180}	r_{180}	r_{270}	r_{360}	r_{90}	v	h	d_2	d_1
r_{270}	r_{270}	r_{360}	r_{90}	r_{180}	d_2	d_1	h	v
h	h	d_2	v	d_1	r_{360}	r_{180}	r_{270}	r_{90}
v	v	d_1	h	d_2	r_{180}	r_{360}	r_{90}	r_{270}
d_1	d_1	h	d_2	v	r_{90}	r_{270}	r_{360}	r_{180}
d_2	d_2	v	d_1	h	r_{270}	r_{90}	r_{180}	r_{360}

(2.1)

We leave it for the reader to verify that the set of rigid motions of a square is a group under the operation $*$. *We also note that* r_{360} *is the identity element.*

This group of rigid motions of a square is known as the ***group of symmetries of the square****. Let us denote this group by Sym. Then*

$$Sym = \{r_{360}, r_{90}, r_{180}, r_{270}, h, v, d_1, d_2\}.$$

Now $h * r_{270} = d_1 \neq d_2 = r_{270} * h$. *Therefore, the group Sym is noncommutative.*

Let us now determine the order of the elements. Consider r_{90}. *Now*

$$r_{90}^2 = r_{90} * r_{90} = r_{180},\ r_{90}^3 = r_{90}^2 * r_{90} = r_{270},$$

and

$$r_{90}^4 = r_{90}^3 * r_{90} = r_{360}.$$

Thus, $\circ(r_{90}) = 4$. *Next, we have*

$$\begin{array}{rcl} \circ(r_{180}) & = & \circ(r_{90}^2) \\ & = & \frac{4}{\gcd(4,2)} \quad \textit{(by Theorem 2.1.46)} \\ & = & \frac{4}{2} \\ & = & 2. \end{array}$$

Similarly,

$$\circ(r_{270}) = 4,\ \ \circ(h) = 2, \circ(v) = 2, \circ(d_1) = 2,\ \textit{and}\ \circ(d_2) = 2.$$

Let us write $\alpha = r_{90}$ *and* $\beta = d_2$. *Then*

$$\alpha^2 = r_{180},\ \alpha^3 = r_{270},\ \alpha^4 = r_{360},\ \beta * \alpha = v, \beta * \alpha^2 = d_1,\ \textit{and}\ \beta * \alpha^3 = h.$$

Also, note that

$$\beta * \alpha = \alpha^{-1} * \beta = \alpha^3 * \beta.$$

Thus, we see that

$$Sym = \{e, \alpha, \alpha^2, \alpha^3, \beta, \beta * \alpha, \beta * \alpha^2, \beta * \alpha^3\}.$$

Finally, we make the following observations. Consider r_{90}. *We can think of* r_{90} *as a one-one function of* $\{1, 2, 3, 4\}$ *onto* $\{1, 2, 3, 4\}$ *by defining*

$$r_{90}(1) = 2, r_{90}(2) = 3, r_{90}(3) = 4, r_{90}(4) = 1.$$

In a similar manner, we can consider other rigid motions of the square as one-one functions of $\{1, 2, 3, 4\}$ *onto* $\{1, 2, 3, 4\}$.

Remark 2.1.50 *A fundamental phenomenon of nature is that of symmetry. A figure or an object is said to have a symmetry if a rotation, a translation, an inversion, a minor reflection, or a combination of these operations leaves the figure or object indistinguishable from its original position. The 1890s saw the first application of group theory to the natural and physical sciences. An important application of group theory was to crystallography. Groups were used to give a theoretical classification of the different kinds of symmetry arrangements possible within crystalline matter 20 years before experimental means were available for analyzing the crystals themselves.*

Remark 2.1.51 *Group theory is used in quantum mechanics. It is used to study the atom's internal structure. In the 1950s, a new generation of particle accelerators produced a variety of subatomic particles. Group theory was used to predict the existence of a tenth nucleon in a tenfold symmetry scheme of nucleons of which nine particles had already been detected. In 1964, the tracks of Omega-Minus, the tenth nucleon, were identified.*

Worked-Out Exercises

♢ **Exercise 1** Let $G = \{a \in \mathbb{R} \mid -1 < a < 1\}$. Define a binary operation $*$ on G by

$$a * b = \frac{a+b}{1+ab}$$

for all $a, b \in G$. Show that $(G, *)$ is a group.

Solution: Note that $-1 < x < 1$ if and only if $x^2 < 1$ for all $x \in \mathbb{R}$.

Let $a, b \in G$. First we show that $a * b \in G$. Now $a^2 < 1$ and $b^2 < 1$. Thus,

$$(1-a^2)(1-b^2) > 0.$$

This implies that

$$1 - a^2 - b^2 + a^2b^2 > 0.$$

Now $(1+ab)^2 - (a+b)^2 = 1 + a^2b^2 + 2ab - a^2 - b^2 - 2ab = 1 - a^2 - b^2 + a^2b^2 > 0$, so

$$(\frac{a+b}{1+ab})^2 < 1.$$

Therefore, $a * b \in G$. Hence, G is closed under $*$. We now show that $*$ is well defined. Let $a, b, c, d \in G$ and $(a, b) = (c, d)$. Then $a = c$ and $b = d$. Thus,

$$a * b = \frac{a+b}{1+ab} = \frac{c+d}{1+cd} = c * d.$$

So $*$ is well defined. To show that $*$ is associative, let $a, b, c \in G$. Now

$$(a * b) * c = \frac{a+b}{1+ab} * c = \frac{\frac{a+b}{1+ab} + c}{1 + (\frac{a+b}{1+ab})c} = \frac{a+b+c+abc}{1+ab+ac+bc}.$$

Similarly,

$$a * (b * c) = \frac{a+b+c+abc}{1+ab+ac+bc}.$$

Therefore, $(a*b)*c = a*(b*c)$, so $*$ is associative. So far, we have shown that $(G, *)$ is a semigroup.

Now $0 \in G$ and

$$0 * a = \frac{0+a}{1+0a} = a \quad \text{for all } a \in G.$$

This shows that $(G, *)$ satisfies (i) of Theorem 2.1.30.

Let $a \in G$. Then $-a \in G$ and

$$(-a) * a = \frac{-a+a}{1+(-a)a} = 0.$$

Thus, $(G, *)$ satisfies (ii) of Theorem 2.1.30. Consequently, by Theorem 2.1.30, $(G, *)$ is a group.

♢ **Exercise 2** Let $G = \{(a, b) \mid a, b \in \mathbb{R},\ a \neq 0\} = \mathbb{R}\backslash\{0\}\times\mathbb{R}$. Define a binary operation $*$ on G by

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

for all (a, b), $(c, d) \in G$. Show that

(a) $(G, *)$ is a group,

(b) G has exactly one element of order 2,

(c) G has no elements of order 3.

Solution: (a) As in Worked-Out Exercise 1, we show that $(G, *)$ satisfies the conditions of Theorem 2.1.30. Let $(a, b), (c, d) \in G$. Then $a \neq 0$ and $c \neq 0$, so $ac \neq 0$. Thus, $(a, b) * (c, d) = (ac, b + d) \in G$. Hence, G is closed under $*$.

It is a direct computation to verify that $*$ is well defined and associative, so we ask the reader to do the verification. Now $(1, 0) \in G$ and

$$(1, 0) * (a, b) = (1a, 0 + b) = (a, b) \quad \text{for all } (a, b) \in G.$$

This implies that $(G, *)$ satisfies (i) of Theorem 2.1.30.

Let $(a, b) \in G$. Then $a \neq 0$, so $\frac{1}{a} \in \mathbb{R}$ and $\frac{1}{a} \neq 0$. Thus, $(\frac{1}{a}, -b) \in G$. Now

$$(\frac{1}{a}, -b) * (a, b) = (\frac{1}{a}a, -b + b) = (1, 0).$$

This implies that $(G, *)$ satisfies (ii) of Theorem 2.1.30.

By Theorem 2.1.30, we can conclude that $(G, *)$ is a group.

(b) First note that $(-1, 0) \in G$ and

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

Thus, $(-1, 0)$ is of order 2. We now show that this is the only element of order 2 by showing that if (a, b) is any other element of G of order 2, then $(a, b) = (-1, 0)$.

Let $(a, b) \in G$ be an element of order 2. Then

$$(a, b) * (a, b) = (1, 0)$$

implies that

$$(a^2, b + b) = (1, 0).$$

Therefore, $a^2 = 1$ and $b = 0$. Now $a^2 = 1$ implies that $a = \pm 1$. If $a = 1$, then $(a, b) = (1, 0)$, which is a contradiction because $(1, 0)$ is of order 1. Hence, $a = -1$. This implies that $(a, b) = (-1, 0)$. It now follows that $(-1, 0)$ is the only element of order 2.

(c) Suppose that (a, b) is an element of order 3. Then

$$(a, b) * (a, b) * (a, b) = (1, 0).$$

This implies that

$$(a^3, 3b) = (1, 0).$$

Thus, $a^3 = 1$ and $b = 0$. Now $a^3 = 1$ implies that $a = 1$. Hence, $(a, b) = (1, 0)$. This means that $(1, 0)$ is of order 2. However, $(1, 0)$ is of order 1, so we have a contradiction. Consequently, G has no element of order 3.

♢ **Exercise 3** Let G be the set of all rational numbers except -1. Show that $(G, *)$ is a group where

$$a * b = a + b + ab$$

for all $a, b \in G$.

Solution: As in Worked-Out Exercise 1, we show that $(G, *)$ satisfies the conditions of Theorem 2.1.30. Our first step is to show that $*$ is well defined. Let $a, b, c, d \in G$ and $(a, b) = (c, d)$. Then $a = c$ and $b = d$. Thus,

$$a * b = a + b + ab = c + d + cd = c * d.$$

Hence, $*$ is well defined. Let $a, b \in G$. Then $a \neq -1$ and $b \neq -1$. We now show that $a * b \in G$ by showing that $a * b \neq -1$ and $a * b$ is a rational number. Suppose $a * b = -1$, i.e., $a + b + ab = -1$. Then

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

This implies that either $(a + 1) = 0$ or $(b + 1) = 0$, i.e., either $a = -1$ or $b = -1$, which is a contradiction. Therefore, $a * b \neq -1$.

Now the addition and multiplication of rational numbers is a rational number, so it follows that $a * b$ is a rational number. Hence, $a * b \in G$. Thus, $*$ is a binary operation on G.

Let $a, b, c \in G$. Then

$$\begin{aligned}(a * b) * c &= (a + b + ab) * c \\ &= a + b + ab + c + ac + bc + abc \\ &= a + (b + c + bc) + a(b + c + bc) \\ &= a + b * c + a(b * c) \\ &= a * (b * c).\end{aligned}$$

This shows that $*$ is associative. Thus, $(G, *)$ is a semigroup. Now $0 \in G$ and

$$0 * a = 0 + a + 0 \cdot a = a$$

for all $a \in G$. Hence, $(G, *)$ satisfies (i) of Theorem 2.1.30. Now for all $a \in G$, $a + 1 \neq 0$. Note that $-\frac{a}{a+1} \neq -1$. Therefore, $-\frac{a}{a+1} \in G$ and

$$-\frac{a}{a+1} * a = -\frac{a}{a+1} + a + (-\frac{a}{a+1})a = \frac{-a + a + a^2 - a^2}{a+1} = 0.$$

This implies that $(G, *)$ satisfies (ii) of Theorem 2.1.30. Hence, by Theorem 2.1.30, $(G, *)$ is a group.

♢ **Exercise 4** Let G be a group and $x \in G$. Suppose $\circ(x) = mn$, where m and n are relatively prime. Show that there exist y, $z \in G$ such that $x = y * z = z * y$ and $\circ(y) = m$ and $\circ(z) = n$.

Solution: Because $\gcd(m, n) = 1$ there exist $s, t \in \mathbb{Z}$ such that $1 = ms + nt$. Now $x = x^{ms+nt} = x^{ms} * x^{nt}$. Let $y = x^{nt}$ and $z = x^{ms}$. Then $x = y * z = z * y$. Now $y^m = (x^{nt})^m = x^{mnt} = e$. Hence, $\circ(y)$ divides m. Similarly, $\circ(z)$ divides n. Suppose $\circ(y) = m_1$ and $\circ(z) = n_1$. It is an easy exercise to verify that $(y * z)^l = y^l * z^l$ for all positive integers l. Thus, $x^{m_1 n_1} = (y * z)^{m_1 n_1} = y^{m_1 n_1} * z^{m_1 n_1} = e * e = e$. Hence, $mn \mid m_1 n_1$. But because $m_1 \mid m$ and $n_1 \mid n$, we must have $m = m_1$ and $n = n_1$.

♢ **Exercise 5** Let $(G, *)$ be a group of even order. Show that there exists $a \in G$ such that $a \neq e$, $a^2 = e$.

Solution: Let $A = \{g \in G \mid g \neq g^{-1}\} \subseteq G$. Then $e \notin A$. If $g \in A$, then $g^{-1} \in A$, i.e., elements of A occurs in pairs. Therefore, the number of elements in A is even. This implies that the number of elements in $\{e\} \cup A$ is odd. Because the number of elements in G is even and $\{e\} \cup A \subseteq G$, there exists $a \in G$ such that $a \notin \{e\} \cup A$. But then $a \neq e$ and $a \notin A$. Hence, there exists $a \in G$ such that $a \neq e$ and $a = a^{-1}$ or $a^2 = e$.

♢ **Exercise 6** Let $(G, *)$ be a group and $a, b \in G$. Suppose that $a * b = b * a^{-1}$ and $b * a = a * b^{-1}$. Show that $a^4 = b^4 = e$.

Solution: Because $a * b = b * a^{-1}$, $a = b * a^{-1} * b^{-1}$. Similarly, $b = a * b^{-1} * a^{-1}$. Thus, $b * a = a * b^{-1} = (b * a^{-1} * b^{-1}) * b^{-1} = b * a^{-1} * b^{-2}$. Multiply both sides of the equation $b * a = b * a^{-1} * b^{-2}$ by b^{-1} to get $a = a^{-1} * b^{-2}$. This implies that $a^2 = b^{-2}$. Hence, $a^4 = a^2 * a^2 = a^2 * b^{-2} = a * (a * b^{-1}) * b^{-1} = a * (b * a) * b^{-1} = (a * b) * a * b^{-1} = (b * a^{-1}) * a * b^{-1} = b * (a^{-1} * a) * b^{-1} = b * e * b^{-1} = e$. Also, $b^4 = a^{-4} = e$.

Exercise 7 Let $(G, *)$ be a group and $a, b \in G$. Suppose that $a * b^n = b^{n+1} * a$ and $b * a^n = a^{n+1} * b$ for some $n \in \mathbb{N}$. Show that $a = b = e$.

Solution: Multiply both sides of the equation $a * b^n = b^{n+1} * a$ by b^{-n} to get $a = b^{n+1} * a * b^{-n}$. Thus, $a^2 = a*a = a* b^{n+1}*a*b^{-n} = (a* b^n)*b*a*b^{-n} = (b^{n+1}*a)*b*a*b^{-n} = b^{n+1}*(a*b)*a*b^{-n}$. Now $a^3 = a*a^2 = a*(b^{n+1}*(a*b)*a*b^{-n}) = (a*b^n)*b*(a*b)*a*b^{-n} = (b^{n+1}*a)*b*(a*b)*a*b^{-n} = b^{n+1} * (a * b)^2 * a * b^{-n}$. Hence, we see that we could use induction to obtain

$$a^n = b^{n+1} * (a * b)^{n-1} * a * b^{-n} \tag{2.2}$$

for all $n \in \mathbb{N}$. Also,

$$\begin{aligned} b * a^n &= a^{n+1} * b \\ &= a * a^n * b \\ &= a * (b^{n+1} * (a * b)^{n-1} * a * b^{-n}) * b \\ &= a * b^{n+1} * (a * b)^{n-1} * a * b^{1-n} \\ &= (a * b^n) * b * (a * b)^{n-1} * a * b^{1-n} \\ &= (b^{n+1} * a) * b * (a * b)^{n-1} * a * b^{1-n} \\ &= b^{n+1} * (a * b)^n * a * b^{1-n}, \end{aligned}$$

which implies that

$$a^n = b^n * (a * b)^n * a * b^{1-n}. \tag{2.3}$$

From Eqs. (2.2) and (2.3),

$$b^{n+1} * (a * b)^{n-1} * a * b^{-n} = b^n * (a * b)^n * a * b^{1-n},$$

which implies that

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

Thus,

$$\begin{aligned} (a * b)^{n+1} &= b * (a * b)^{n-1} * a \\ &= b * \underbrace{((a * b) * \cdots * (a * b))}_{n-1 \text{ times}} * a \\ &= \underbrace{(b * a) * \cdots * (b * a)}_{n \text{ times}} \\ &= (b * a)^n. \end{aligned} \tag{2.4}$$

Interchange the role of a and b to get

$$(b * a)^{n+1} = (a * b)^n. \tag{2.5}$$

Hence, $(a * b)^n = (b * a)^{n+1} = (b * a)^n * (b * a) = (a * b)^{n+1} * (b * a)$, so $e = (a * b) * (b * a)$, which implies that

$$a^2 = b^{-2}. \tag{2.6}$$

Now

$$b * a^n = b * a^2 * a^{n-2} = b * b^{-2} * a^{n-2} = b^{-1} * a^{n-2} \tag{2.7}$$

and

$$a^{n+1} * b = a^{n-1} * a^2 * b = a^{n-1} * b^{-2} * b = a^{n-1} * b^{-1}. \tag{2.8}$$

Thus, from Eqs. (2.7) and (2.8) it follows that $b^{-1} * a^{n-2} = a^{n-1} * b^{-1}$, so

$$a^{n-1} = b^{-1} * a^{n-2} * b = (b^{-1} * a * b)^{n-2}. \tag{2.9}$$

Now $b * a^n = a^{n+1} * b$ implies that

$$a^n = (b^{-1} * a * b)^{n+1}. \tag{2.10}$$

Hence, $a^n = (b^{-1} * a * b)^{n+1} = (b^{-1} * a * b)^{n-2} * (b^{-1} * a * b)^3 = a^{n-1} * (b^{-1} * a * b)^3$, which implies that $a = (b^{-1} * a * b)^3 = b^{-1} * a^3 * b$. Thus, $a^3 * b = b * a$. Therefore, $b * a = a^3 * b = a* a^2 * b = a * b^{-2} * b = a * b^{-1}$ by Eq. (2.6). That is, we have

$$b * a = a * b^{-1}. \tag{2.11}$$

Similarly,

$$a * b = b * a^{-1}. \tag{2.12}$$

Now $a*b = b*a^{-1}$ implies that $a*b*a = b$. Thus, $b = a*b*a = a*a*b^{-1}$ [by Eq. (2.11)]. Hence,

$$a^2 = b^2.$$

Suppose n is even. Then $a^2 = b^2$ implies that $a^n = b^n$. Hence, $a * b^n = b^{n+1} * a$ implies that $a^{n+1} = a^n * b * a$, so $b = e$. Similarly, $a = e$. Suppose n is odd. Let $n = 2k + 1$. Then $a^{2k} = b^{2k}$. Now $a * b^n = b^{n+1} * a \Rightarrow a * b^{2k+1} = b^{2k+2} * a \Rightarrow a * a^{2k} * b = a^{2k+2} * a$. Thus, $b = a^2 = b^2$. Hence, $b = e$. Similarly, $a = e$.

Exercise 8 (**Hays**) Let $(S, *)$ be a semigroup. Show that S is a group if and only if for all $a \in S$ there exists a unique $b \in S$ such that $a * b * a = a$.

Solution: Suppose for all $a \in S$, there exists a unique $b \in S$ such that $a * b * a = a$. Let $a \in S$. Then there exists $b \in S$ such that $a * b * a = a$. Thus, $a * b * a * b = a * b$, so $(a * b)^2 = a * b$. Hence, S has an idempotent element. If $(S, *)$ is to be a group, then it can have only one idempotent (Corollary 2.1.17), namely, the identity element. Therefore, first we show that S has only one idempotent.

Suppose e and f are two idempotents in S. Because $e * f \in S$, there exists a unique g such that $(e * f) * g * (e * f) = e * f$. Now $(e * f) * (g * e) * (e * f) = (e * f) * g * (e * e) * f = (e*f)*g*e^2*f = (e*f)*g*(e*f) = e*f$. Because g is unique such that $(e*f)*g*(e*f) = (e*f)$, it follows that $g*e = g$. Similarly, because $(e*f)*(f*g)*(e*f) = (e*f)*g*(e*f) = e*f$, the uniqueness of g implies that $f * g = g$. Also, $(e * f) * (g * (e * f) * g) * (e * f) = ((e * f) * g * (e * f)) * g * (e * f) = (e * f) * g * (e * f)$. Again, the uniqueness of g implies that $g * (e * f) * g = g$. Hence, $g^2 = g * g = (g*e)*(f*g) = g*(e*f)*g = g$. Thus, g is an idempotent. Now $g = g*g*g$ and $g*(e*f)*g = g$. Hence, by the uniqueness of the middle element $g = e * f$. Therefore, $e * f$ is an idempotent. Now $(e*f)*f*(e*f) = (e*(f*f))*(e*f) = (e*f)*(e*f) = e*f$ and similarly $(e*f)*e*(e*f) = e*f$. By the uniqueness of the middle element, it follows that $e = f$. Hence, S has a unique idempotent element.

Let e be the idempotent element of S. Let $a \in S$. Then there exists $b \in S$ such that $a * b * a = a$, which implies that $(a*b)^2 = a*b$. Hence, $a*b = e$. Also, $a*b*a = a$ implies that $b*a*b*a = b*a$. Thus, $b * a$ is an idempotent. Hence, $b * a = e$. Also, $a * b * a = a$ together with $a * b = e = b * a$ implies that $e * a = a = a * e$. Therefore, e is the identity element. Because $a * b = e = b * a$, b is an inverse of a. Consequently, $(S, *)$ is a group.

Conversely, suppose $(S, *)$ is a group. Let $a \in S$. Note that $a * a^{-1} * a = a$. This shows the existence of an element $b \in S$ such that $a * b * a = a$, namely, $b = a^{-1}$. To show the uniqueness, suppose there exist $b, c \in S$ such that $a * b * a = a$ and $a * c * a = a$. Then $a * b * a = a * c * a$ and by the cancellation laws, $b = c$. Thus, b is unique such that $a * b * a = a$.

Exercises

1. Which of the following mathematical systems are semigroups? Which are groups?

   (a) $(\mathbb{N}, *)$, where $a * b = a$ for all $a, b \in \mathbb{N}$.

   (b) $(\mathbb{Z}, *)$, where $a * b = a - b$ for all $a, b \in \mathbb{Z}$.

   (c) $(\mathbb{R}, *)$, where $a * b = |a|b$ for all $a, b \in \mathbb{R}$.

   (d) $(\mathbb{R}, *)$, where $a * b = a + b + 1$ for all $a, b \in \mathbb{R}$.

   (e) $(\mathbb{R}, *)$, where $a * b = a + b - ab$ for all $a, b \in \mathbb{R}$.

   (f) $(\mathbb{Q}, *)$, where $a * b = \frac{ab}{2}$ for all $a, b \in \mathbb{Q}$.

   (g) $(G, *)$, where

   $$G = \left\{ \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \mid \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \neq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } a, b \in \mathbb{R} \right\}$$

   and $*$ is the usual matrix multiplication.

(h) $(G, *)$, where G is the set of all matrices of the following form over $\mathbb{Z}$

$$\begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix}$$

and $*$ is the usual matrix multiplication.

2. Let $G = \{(a, b) \mid a, b \in \mathbb{R},\ b \neq 0\}$. Define a binary operation $*$ on G by $(a, b) * (c, d) = (a + bc, bd)$ for all $(a, b), (c, d) \in G$. Show that $(G, *)$ is a noncommutative group.

3. Let $G = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a, b, c, d \in \mathbb{R},\ ad - bc = 1 \right\}$. Show that G is a group under usual matrix multiplication. (This group is usually denoted by $SL(2, \mathbb{R})$ and is called the **special linear group of degree 2**.)

4. Let $G = \left\{ \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix} \mid n \in \mathbb{Z} \right\}$. Show that $(G, *)$ is a commutative group, where $*$ denotes the usual matrix multiplication. Also, show that $(G, *)$ is torsion-free.

5. In $\mathbb{Z}_{14}$, find the smallest positive integer n such that $n[6] = [0]$.

6. Find an element $[b] \in \mathbb{Z}_9$ such that $[8] \cdot_9 [b] = [1]$. Does $[b] \in U_9$?

7. In U_{24}, find the smallest positive integer n such that $[7]^n = [1]$.

8. Describe U_6, U_9, U_{12}, U_{24} of Example 2.1.10.

9. Let p be a prime. Show that $U_p = \mathbb{Z}_p \backslash \{[0]\}$.

10. Let $U_n = \{[a] \in \mathbb{Z}_n \backslash \{[0]\} \mid \gcd(a, n) = 1\}$. Show that $(U_n, \cdot_n)$ is a group, where $\cdot_n$ is the multiplication modulo n.

11. Show that $U_n = \{[a] \in \mathbb{Z}_n \backslash \{[0]\} \mid \text{additive order of } [a] = n\ \}$.

12. Let $(G, *)$ be a group and $a, b \in G$. Suppose that $a^2 = e$ and $a * b^4 * a = b^7$. Show that $b^{33} = e$.

13. Let $(G, *)$ be a group and $a, b \in G$. Suppose that $a^{-1} * b^2 * a = b^3$ and $b^{-1} * a^2 * b = a^3$. Show that $a = b = e$.

14. Let $(G, *)$ be a group. If $a, b \in G$ are such that $a^4 = e$ and $a^2 * b = b * a$, show that $a = e$.

15. Let $(G, *)$ be a group and $x, a, b \in G$. Let $c = x * a * x^{-1}$ and $d = x * b * x^{-1}$. Show that $a * b = b * a$ if and only if $c * d = d * c$.

16. Let $(G, *)$ be a group such that $a^2 = e$ for all $a \in G$. Show that G is commutative.

17. Prove that a group $(G, *)$ is commutative if and only if $(a * b)^{-1} = a^{-1} * b^{-1}$ for all $a, b \in G$.

18. Let $(G, *)$ be a group. Prove that if $(a * b)^2 = a^2 * b^2$ for all $a, b \in G$, then $(G, *)$ is commutative.

19. Prove that a group $(G, *)$ is commutative if and only if for all $a, b \in G$, $(a * b)^n = a^n * b^n$ for any three consecutive integers n.

20. Let $(G, *)$ be a group. If G has only two elements, prove that G is commutative.

21. Let $(G, *)$ be a group and $a, b, c \in G$. Find an element $x \in G$ such that $a * x * b = c$. Is x unique?

22. Let $(G, *)$ be a group and $a, b \in G$. Show that $(a * b * a^{-1})^n = a * b^n * a^{-1}$ for all integers n.

23. Let $(G, *)$ be a finite group and $a \in G$. Show that there exists $n \in \mathbb{N}$ such that $a^n = e$.

24. If $(G, *)$ is a group and $a_1, \ldots, a_n \in G$, prove that $(a_1 * \cdots * a_n)^{-1} = a_n^{-1} * \cdots * a_1^{-1}$.

25. Let $(G, *)$ and $(H, \cdot)$ be groups. Define the operation $\star$ on $G \times H = \{(a, b) \mid a \in G,\ b \in H\}$ by $(a, b) \star (c, d) = (a * c, b \cdot d)$. Prove that $(G \times H, \star)$ is a group. If $(G, *)$ and $(H, \cdot)$ are commutative, prove that $(G \times H, \star)$ is commutative. The group $(G \times H, \star)$ is called the **direct product** of G and H.

26. Let $(G, *)$ be a finite group and $a \in G$. Show that $\circ(a) \leq |G|$.

27. Let $(G, *)$ be a group and $a, b \in G$.

    (a) Prove that a and a^{-1} have the same order.

    (b) Prove that a and $b * a * b^{-1}$ have the same order.

    (c) Prove that $a * b$ and $b * a$ have the same order.

28. Let $(G, *)$ be a group and $a, b \in G$.

    (a) Suppose that $a * b = b^5 * a^3$. Show that $\circ(b * a^{-1}) = \circ(b^5 * a) = \circ(b^3 * a^3)$.

    (b) Generalize (a) to arbitrary powers of a and b.

29. Let $(G, *)$ be a group, $a \in G$ and $\circ(a) = n$. Let $1 \leq p \leq n$ be such that p and n are relatively prime. Show that $\circ(a^p) = n$.

30. Let $(G, *)$ be a group, $a \in G$, and $\circ(a) = p$, where p is a prime.

    (a) Prove that $\circ(a^k) = p$ for all $1 \leq k < p$.

    (b) Prove that for all $m \in \mathbb{N}$, either $a^m = e$ or $\circ(a^m) = p$.

31. Let $(G, *)$ be a group and $a \in G$. Suppose that $\circ(a) = n$ and $n = mk$ for some $m, k \in \mathbb{N}$. What is $\circ(a^k)$?

32. Let $(G, *)$ be a group.

    (a) Let $a, b \in G$, $\circ(a) = n$, $\circ(b) = m$, $\gcd(m, n) = 1$, and $a * b = b * a$. Show that $\circ(a * b) = mn$.

    (b) Let $a_i \in G$, $\circ(a_i) = n_i$, $1 \leq i \leq m$. Suppose $\gcd(n_i, n_j) = 1$ and $a_i a_j = a_j a_i$ for all i and j. Let $x = a_1 * a_2 * \cdots * a_m$. Show that $\circ(x) = n_1 n_2 \cdots n_m$.

33. Let $(G, *)$ be a group and $x \in G$. Suppose $\circ(x) = n = n_1 n_2 \cdots n_k$, where for all $i \neq j$, n_i and n_j are relatively prime. Show that there exists $x_i \in G$ such that $\circ(x_i) = n_i$ for all $i = 1, 2, \ldots, k$, $x = x_1 * x_2 * \cdots * x_k$ and $x_i * x_j = x_j * x_i$ for all i and j.

34. Let $G = \{(a, b) \mid a, b \in \mathbb{R},\ a \neq 0\}$. Then G is a group under the binary operation $(a, b) * (c, d) = (ac, bc + d)$ for all $(a, b), (c, d) \in G$. Prove that G has infinitely many elements of order 2, but G has no element of order 3.

35. Let $a, b \in \text{Sym}$. As remarked in Example 2.1.49, every rigid motion of the square can be considered a one-one function of $\{1, 2, 3, 4\}$ onto itself. Consider $a * b$ as a function. Show that $a * b = a \circ b$, where $*$ represents the binary operation of rigid motions of the square and $\circ$ is the composition of functions.

36. Let $(S, *)$ be a finite semigroup. Prove that there exists $a \in S$ such that $a^2 = a$.

37. Let $(G, *)$ be a finite semigroup with identity. Prove that $(G, *)$ is a group if and only if G has only one element a such that $a^2 = a$.

38. Prove that a semigroup $(S, *)$ is a group if and only if $a * S = S$ and $S * a = S$ for all $a \in S$, where $a * S = \{a * s \mid s \in S\}$ and $S * a = \{s * a \mid s \in S\}$.

39. Prove that a semigroup $(S, *)$ is a group if and only if

(a) there exists $e \in S$ such that $a * e = a$ for all $a \in S$, and

(b) for all $a \in S$ there exists $b \in S$ such that $a * b = e$.

40. Rewrite the statements and proofs of the theorems in this chapter using additive notation.

41. Let $(G, *)$ be a group, $a, b \in G$ and $m, n \in \mathbb{Z}$. Prove that

(a) $a^n * a^m = a^{n+m} = a^m * a^n$,

(b) $(a^n)^m = a^{nm}$,

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

(d) $e^n = e$,

(e) $(a * b)^n = a^n * b^n$, if $(G, *)$ is commutative.

42. Write the proof if the following statements are true; otherwise, give a counterexample.

(a) Let $T(S)$ be the set of all functions on $S = \{1, 2, 3\}$. $T(S)$ is a group under composition of functions.

(b) $M_2(\mathbb{R}) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a, b, c, d \in \mathbb{R} \right\}$ is a group under usual matrix multiplication.

(c) Every group of four elements is commutative.

(d) A group has only one idempotent element.

(e) A semigroup with only one idempotent is a group.

(f) If a semigroup S satisfies the cancellation laws, then S is a group.

Niels Henrik Abel (1802–1829) was born on August 5, 1802, in Finnöy, Norway. He was the second of six children. Abel and his brothers received their first education from their father.

At the age of 13, Abel along with his older brother, was sent to the Cathedral sch-ool in Christiania (Oslo). In 1817, his mathematics teacher was Bernt Michael Holmbë, who was seven years older than Abel. Holmbë recognized Abel's talent and started giving him special problems and recommended special books outside the curriculum. Abel and Holmbë read the calculus text of Euler and the work of Lagrange and Laplace. Soon Abel became familiar with most of the important mathematical literature.

Abel's father died when he was 18 years old and the responsibility of supporting the family fell on his shoulders. He gave private lessons and did odd jobs. However, he continued to carry out his mathematical research.

Abel, in his last year of school, attacked the problem of the solvability of the quintic equation, a problem that had been unsettled since the sixteenth century. Abel thought that he had solved the problem and submitted his work for publication. Unable to find an error and understand his arguments, he was asked by the editor to illustrate his method. In 1824, during the process of illustration he discovered an error. This discovery led Abel to a proof that no such solution exists. He also worked on elliptic functions and in essence revolutionized the theory of elliptic functions.

He traveled to Paris and Berlin in order to find a teaching position. Then poverty took its toll, and Abel died from tuberculosis on April 6, 1829. Two days later a letter from Crelle reached his address, conveying the news of his appointment to the professorship of mathematics at the University of Berlin.

Abel is honored by such terms as Abelian group and Abelian function.

Chapter 3

Permutation Groups

Permutation groups is one of the specialized theories of groups which arose from the source, classical algebra, in the evolution of group theory.

3.1 Permutation Groups

As stated earlier, there are four major sources from which abstract group theory evolved. Mathematicians' interest in finding formulas to solve polynomial equations by means of radicals led some mathematicians to the study of permutations of the roots of rational functions. Lagrange, Rufini, and Cauchy were among the earlier mathematicians to work with permutation groups. However, it was Cauchy whose systematic study of permutation groups (between 1815 and 1845) is believed, by some, to be the origin of abstract group theory. Many of the concepts and major results in this chapter are due to Cauchy.

We begin our study of permutation groups by defining what a permutation is.

Definition 3.1.1 *Let X be a nonempty set. A **permutation** π of X is a one-one function from X onto X.*

Example 3.1.2 *(i) Let X be a nonempty set. Define $\pi : X \to X$ by $\pi(x) = x$ for all $x \in X$. Then π is one-one function of X onto X. Thus, π is a permutation of X. Nothe that π is called the identity permutations and is, usually, denoted by i_X or e.*

(ii) Let $X = \{a, b, c\}$. Define $\alpha : X \to X$ such that $\alpha(a) = b$, $\alpha(b) = a$, and $\alpha(c) = c$. By the definition of α it follows that is α is one-one function of X onto X. Thus, α is a permutation of X.

(iii) Consider $\mathbb{R}$, the set of real numbers. Define $\alpha : \mathbb{R} \to \mathbb{R}$ by $\alpha(x) = 3x + 5$ for all $x \in \mathbb{R}$. It can be shown that α is a one-one function of $\mathbb{R}$ onto $\mathbb{R}$. Thus, α is a permutation of $\mathbb{R}$. Similarly, if $\beta : \mathbb{R} \to \mathbb{R}$ by $\beta(x) = x^3$ for all $x \in \mathbb{R}$. It can be shown that β is a one-one function of $\mathbb{R}$ onto $\mathbb{R}$. Thus, β is a permutation of $\mathbb{R}$.

Definition 3.1.3 *A group $(G, *)$ is called a **permutation** group on a nonempty set X if the elements of G are permutations of X and the operation $*$ is the composition of two functions.*

Example 3.1.4 *Let X be any nonempty set and S_X be the set of all one-one functions from X onto X, as defined in Example* 2.1.13. *Then $(S_X, \circ)$ is a group as we have shown in Example* 2.1.13, *where $\circ$ is the composition of functions. Hence, $(S_X, \circ)$ is a permutation group.*

Example 3.1.5 *Let $X = \{1, 2\}$. Define $\alpha : X \to X$ such that $\alpha(1) = 1$, $\alpha(2) = 2$. Then α is a one-one function of X onto X, so β is a permutation of X. Next define $\beta : X \to X$ such that $\beta(1) = 2$ and $\beta(2) = 1$. Then β is a one-one function of X onto X, so β is a permutation of X. Let $S_X = \{\alpha, \beta\}$. Then $(S_X, \circ)$ is a group, where $\circ$ is the composition of functions. Note that on this set X, α and b are the only permutations on X. Moreover, α is the identity permutation and $\beta^{-1} = \beta$.*

In this chapter, and in fact in this text, our study of permutation groups will focus on permutation groups on finite sets, i.e., X is a finite set.

Before we consider more examples of permutation groups, let us fix some notation which will be useful when working with permutations.

Let $I_n = \{1, 2, \ldots, n\}$, $n \geq 1$. Let π be a permutation on I_n. Then

$$\pi = \{(1, \pi(1)), (2, \pi(2)), \ldots, (n, \pi(n))\}.$$

(Recall that a function $f : A \to A$ is a subset of $A \times A$.) It is sometimes convenient to describe a permutation by means of the following notational device:

$$\pi = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ \pi(1) & \pi(2) & \pi(3) & \cdots & \pi(n) \end{pmatrix}.$$

This notation is due to Cauchy and is called the **two-row notation**. In the upper row, we list all the elements of I_n and in the lower row under each element $i \in I_n$, we write the image of the element, i.e., $\pi(i)$.

Example 3.1.6 *Let $n = 4$ and π be the permutation on I_4 defined by $\pi(1) = 2$, $\pi(2) = 4$, $\pi(3) = 3$, and $\pi(4) = 1$. Then using the two-row notation we can write*

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

As we shall see, the two-row notation of permutations is quite convenient while doing computations such as determining the composition of permutations.

Let $n = 7$ and π and σ be two permutations on I_7 defined by

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

and

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

Let us compute $\pi \circ \sigma$. Now by the definition of the composition of functions

$$(\pi \circ \sigma)(i) = \pi(\sigma(i))$$

for all $i \in I_7$. Thus,

$$(\pi \circ \sigma)(1) = \pi(\sigma(1)) = \pi(2) = 3,$$
$$(\pi \circ \sigma)(2) = \pi(\sigma(2)) = \pi(5) = 7,$$

and so on. From this, it is clear that when determining, say, $(\pi \circ \sigma)(1)$, we start with σ and finish with π and read as follows: 1 goes to 2 (under σ) and 2 goes to 3 (under π), so 1 goes to 3 (under $\pi \circ \sigma$). We can exhibit this in the following form:

$$\begin{array}{ll} 1 \xrightarrow{\sigma} 2 \xrightarrow{\pi} 3 & 1 \overset{\pi\circ\sigma}{\to} 3 \\ 2 \xrightarrow{\sigma} 5 \xrightarrow{\pi} 7 & 2 \overset{\pi\circ\sigma}{\to} 7 \\ 3 \xrightarrow{\sigma} 3 \xrightarrow{\pi} 4 & 3 \overset{\pi\circ\sigma}{\to} 4 \\ 4 \xrightarrow{\sigma} 1 \xrightarrow{\pi} 1 & 4 \overset{\pi\circ\sigma}{\to} 1 \\ 5 \xrightarrow{\sigma} 7 \xrightarrow{\pi} 5 & 5 \overset{\pi\circ\sigma}{\to} 5 \\ 6 \xrightarrow{\sigma} 6 \xrightarrow{\pi} 2 & 6 \overset{\pi\circ\sigma}{\to} 2 \\ 7 \xrightarrow{\sigma} 4 \xrightarrow{\pi} 6 & 7 \overset{\pi\circ\sigma}{\to} 6. \end{array}$$

Thus,

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

Example 3.1.7 *Let $n = 6$ and α and β be permutations on I_6 defined by*

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 4 & 6 & 5 & 2 \end{pmatrix}$$

and

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

Let us first determine $\alpha \circ \beta$. Now

$$1 \xrightarrow{\beta} 1 \xrightarrow{\alpha} 3, \text{ i.e., } 1 \xrightarrow{\alpha\circ\beta} 3.$$

Similarly,

$$2 \xrightarrow{\alpha\circ\beta} 4, 3 \xrightarrow{\alpha\circ\beta} 5, 4 \xrightarrow{\alpha\circ\beta} 6, 5 \xrightarrow{\alpha\circ\beta} 1, 6 \xrightarrow{\alpha\circ\beta} 2.$$

Thus,

$$\alpha \circ \beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 6 & 1 & 2 \end{pmatrix}.$$

Similarly, for $\beta \circ \alpha$; $1 \xrightarrow{\alpha} 3 \xrightarrow{\beta} 5$, i.e., $1 \xrightarrow{\beta\circ\alpha} 5$, and so on. In this case, we start with α and finish with β. Note that

$$\beta \circ \alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 5 & 1 & 4 & 6 & 2 & 3 \end{pmatrix}.$$

We note that $\alpha \circ \beta \neq \beta \circ \alpha$.

Let S_n denote the set of all permutations on I_n, $n \geq 1$.

Example 3.1.8 *In this example, we describe S_3, i.e., the set of all permutations on $I_3 = \{1, 2, 3\}$. From Exercise 8 (page 31), we know that the number of one-one functions of I_3 onto I_3 is $3! = 6$. Thus, $|S_3| = 6$. Let e denote the identity permutation on I_3, i.e., $e = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}$. Let α_1 be a nonidentity permutation on I_3. Let us see some of the choices for α_1. Suppose $\alpha_1(1) = 1$. If $\alpha_1(2) = 2$, then we must have $\alpha_1(3) = 3$ because α_1 is a permutation. In this case, we see that $\alpha_1 = e$, a contradiction. Thus, we must have $\alpha_1(2) = 3$ and $\alpha_1(3) = 2$, i.e., $\alpha_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$. In a similar manner, we can show that the other four permutations on I_3 are $\alpha_2 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$, $\alpha_3 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$, $\alpha_4 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$, and $\alpha_5 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$. Thus,*

$$S_3 = \{e, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}.$$

Let us denote α_2 by α and α_4 by β. We ask the reader to check that

$$\beta^2 = \alpha_5, \alpha \circ \beta = \alpha_1, \text{ and } \alpha \circ \beta^2 = \alpha_3.$$

Hence, we can write

$$S_3 = \{e, \beta, \beta^2, \alpha, \alpha \circ \beta, \alpha \circ \beta^2\}.$$

Because $(S_3, \circ)$ is also a group, we ask the reader to show that $\circ(\alpha) = 2$ and $\circ(\beta) = 3$ by showing that $\alpha^2 = e$ and $\beta^2 \neq e$, but $\beta^3 = e$.

In the previous example, the permutation group $(S_3, \circ)$ consisted of all the permutations on the set I_3. Next, we give an example of a permutation group that does not contain all the permutations on a given set.

Example 3.1.9 *Let $n = 4$ and consider $I_4 = \{1, 2, 3, 4\}$. Recall that in Example* 2.1.49*, we remarked that rigid motions of the square can be viewed as permutations on I_4. Let S be the set of all permutations that corresponds to the rigid motions of the square. We will use the same notation for the permutations, i.e., r_{90} is the permutation* $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}$*, r_{360} is the identity permutation, and so on. By Exercise* 35 *(page* 56*), it follows that the multiplication table of $(S, \circ)$ is the same as the multiplication table of the group (Sym, $*$). Now composition of functions is associative and from the multiplication table, it follows that S is closed under $\circ$, r_{360} is the identity of $(S, \circ)$, and every element of S has an inverse. Thus, $(S, \circ)$ is a group. Hence, the group of symmetries of a square can be thought of as a permutation group on I_4.*

The following theorem describes some basic properties of S_n.

Theorem 3.1.10 *(i) $(S_n, \circ)$ is a group for any positive integer $n \geq 1$.*
(ii) If $n \geq 3$, then $(S_n, \circ)$ is noncommutative.
(iii) $|S_n| = n!$

Proof. (i) We have already noted that the set of all one-one functions of any nonempty set onto itself forms a group under composition of functions in Example 2.1.13. Thus, $(S_n, \circ)$ is a group for any positive integer $n \geq 1$.

(ii) Let $n \geq 3$. Let $\alpha, \beta \in S_n$ be defined by

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n \\ 1 & 3 & 2 & 4 & \cdots & n \end{pmatrix} \text{ and } \beta = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n \\ 3 & 2 & 1 & 4 & \cdots & n \end{pmatrix}.$$

Now

$$\alpha \circ \beta = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n \\ 2 & 3 & 1 & 4 & \cdots & n \end{pmatrix}$$

and

$$\beta \circ \alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n \\ 3 & 1 & 2 & 4 & \cdots & n \end{pmatrix}.$$

Thus, $(\alpha \circ \beta)(1) = 2 \neq 3 = (\beta \circ \alpha)(1)$. Hence, $\alpha \circ \beta \neq \beta \circ \alpha$, so S_n is noncommutative.

(iii) This follows from Exercise 8 (page 31). ■

Definition 3.1.11 *The group $(S_n, \circ)$ is called the* ***symmetric group on*** *I_n.*

Consider the permutation $\pi = \begin{pmatrix} 1 & 2 & \cdots & n \\ \pi(1) & \pi(2) & \cdots & \pi(n) \end{pmatrix}$. If $\pi(i) = i$, then we drop the column $\begin{matrix} i \\ \pi(i) \end{matrix}$. For example, $\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix}$ is denoted by $\begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix}$.

Definition 3.1.12 *Let π be an element of S_n. Then π is called a k-* ***cycle****, written $(i_1\ i_2\ \cdots\ i_k)$, if*

$$\pi = \begin{pmatrix} i_1 & i_2 & \cdots & i_{k-1} & i_k \\ i_2 & i_3 & \cdots & i_k & i_1 \end{pmatrix},$$

i.e., $\pi(i_j) = i_{j+1}$, $j = 1, 2, \ldots, k-1$, $\pi(i_k) = i_1$, and $\pi(a) = a$ for any other element of I_n.

Note that if $\pi = (i_1 i_2 \cdots i_k)$, then

$$\begin{aligned} \pi &= (i_1 i_2 \cdots i_k) \\ &= (i_2 i_3 \cdots i_k i_1) \\ &\vdots \\ &= (i_j i_{j+1} \cdots i_k i_1 \cdots i_{j-1}). \end{aligned}$$

A k-cycle is called a **transposition** when $k = 2$.

We know that in Example 3.1.9, the permutation r_{90} is a 4-cycle and d_2 is a 2-cycle. We write

$$r_{90} = (1\ 2\ 3\ 4)$$

and

$$d_2 = (1\ 3).$$

The identity of S_n is sometimes denoted by (1) or e.

Example 3.1.13 *Using the cycle notation, we can write*

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

We now note some of the properties of the group $(S_3, \circ)$.

(*i*) $(S_3, \circ)$ *is a noncommutative group of order* 6 *by Theorem* 3.1.10.

(*ii*) S_3 *contains two elements of order* 3; *for* $(1\ 2\ 3) \circ (1\ 2\ 3) = (1\ 3\ 2) \neq e$ *and* $(1\ 2\ 3) \circ (1\ 2\ 3) \circ (1\ 2\ 3) = e$. *Hence, the order of* $(1\ 2\ 3)$ *is* 3. *Similarly, the order of* $(1\ 3\ 2)$ *is* 3. *The order of* $(1\ 2), (1\ 3)$, *and* $(2\ 3)$ *is* 2 *because* $(1\ 2) \circ (1\ 2) = e$, $(1\ 3) \circ (1\ 3) = e$, *and* $(2\ 3) \circ (2\ 3) = e$.

(*iii*) *In* S_3, *the product of distinct elements of order* 2 *is an element of order* 3. $(1\ 2) \circ (2\ 3) = (1\ 2\ 3)$, $(1\ 3) \circ (1\ 2) = (1\ 2\ 3)$, $(1\ 2) \circ (1\ 3) = (1\ 3\ 2)$, $(2\ 3) \circ (1\ 2) = (1\ 3\ 2)$, $(1\ 3) \circ (2\ 3) = (1\ 3\ 2)$, *and* $(2\ 3) \circ (1\ 3) = (1\ 2\ 3)$.

Definition 3.1.14 *Let* $\alpha, \beta \in S_n$. *Then* α *and* β *are called* ***conjugate*** *if there exists* $\gamma \in S_n$ *such that*

$$\gamma \circ \alpha \circ \gamma^{-1} = \beta.$$

The following theorem shows how to compute the conjugate of a cycle.

Theorem 3.1.15 *Let* $\pi = (i_1 i_2 \cdots i_l) \in S_n$ *be a cycle. Then for all* $\alpha \in S_n$,

$$\alpha \circ \pi \circ \alpha^{-1} = (\alpha(i_1)\ \alpha(i_2)\ \cdots\ \alpha(i_l)).$$

Proof. Because $\alpha \in S_n$, α is a one-one mapping of I_n onto I_n. Thus, the elements $\alpha(1), \ldots, \alpha(n) \in I_n$ are all distinct, so $I_n = \{\alpha(1), \alpha(2), \ldots, \alpha(n)\}$. Let r be any integer such that $1 \leq r < l$. Then

$$\begin{aligned}(\alpha \circ \pi \circ \alpha^{-1})(\alpha(i_r)) &= \alpha(\pi(\alpha^{-1}(\alpha(i_r)))) \\ &= \alpha(\pi(i_r)) \\ &= \alpha(i_{r+1}).\end{aligned}$$

Also, $(\alpha \circ \pi \circ \alpha^{-1})(\alpha(i_l)) = \alpha(\pi(\alpha^{-1}(\alpha(i_l)))) = \alpha(\pi(i_l)) = \alpha(i_1)$. Now let $a \in I_n$ be such that $a \neq \alpha(i_r)$ for all r, $1 \leq r \leq l$. Then $\alpha^{-1}(a) \in I_n$ and $\alpha^{-1}(a) \neq i_r$ for all r, $1 \leq r \leq l$, so $\pi(\alpha^{-1}(a)) = \alpha^{-1}(a)$. Thus,

$$\begin{aligned}(\alpha \circ \pi \circ \alpha^{-1})(a) &= \alpha(\pi(\alpha^{-1}(a))) \\ &= \alpha(\alpha^{-1}(a)) \\ &= a.\end{aligned}$$

It now follows that $\alpha \circ \pi \circ \alpha^{-1} = (\alpha(i_1)\ \alpha(i_2) \cdots\ \alpha(i_l))$. ■

Definition 3.1.16 *Let* $\pi_1, \pi_2, \ldots, \pi_k \in S_n$. *Then* $\pi_1, \pi_2, \ldots, \pi_k$ *are called* ***disjoint*** *if for all* i, $1 \leq i \leq k$ *and for all* $a \in I_n$, $\pi_i(a) \neq a$ *implies* $\pi_j(a) = a$ *for all* $j \neq i$, $1 \leq j \leq k$.

In other words, $\pi_1, \pi_2, \ldots, \pi_k \in S_n$ are disjoint if for all $1 \leq i \leq k$ and for all $a \in I_n$, if π_i moves a, then all other permutations π_j must fix a, i.e., $\pi_j(a) = a$ for all $j \neq i$, $1 \leq j \leq k$.

Let π and λ be disjoint permutations on I_n. Let $a \in S$ be such that $\pi(a) \neq a$. Then $\lambda(a) = a$. Let $\pi(a) = b$. Then

$$(\pi \circ \lambda)(a) = \pi(\lambda(a)) = \pi(a) = b.$$

Also,

$$(\lambda \circ \pi)(a) = \lambda(\pi(a)) = \lambda(b).$$

If $\pi(b) = b$, then $\pi(b) = b = \pi(a)$, so $a = b$. Thus, $\pi(a) = b = a$, a contradiction. Hence, $\pi(b) \neq b$, so $\lambda(b) = b$. Thus,

$$(\lambda \circ \pi)(a) = \lambda(\pi(a)) = \lambda(b) = b.$$

Hence, $(\pi \circ \lambda)(a) = (\lambda \circ \pi)(a)$. Suppose $\pi(a) = a$. If $\lambda(a) = a$, then $(\pi \circ \lambda)(a) = a = (\lambda \circ \pi)(a)$. Suppose $\lambda(a) \neq a$. By a similar argument as before, $(\pi \circ \lambda)(a) = (\lambda \circ \pi)(a)$. Therefore, $\pi \circ \lambda = \lambda \circ \pi$. Consequently, if π and λ are disjoint permutations, then they commute. We record this result in the following theorem.

Theorem 3.1.17 *Let π, $\lambda \in S_n$ such that π and λ are disjoint. Then $\pi \circ \lambda = \lambda \circ \pi$, i.e., π and λ commute.*

Consider $\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 5 & 1 & 8 & 3 & 7 & 6 & 4 \end{pmatrix} \in S_n$. Then $\pi = (1\ 2\ 5\ 3) \circ (4\ 8) \circ (6\ 7)$ can be written as a product of disjoint cycles. This leads us to the following theorem.

Theorem 3.1.18 *Any nonidentity permutation π of S_n ($n \geq 2$) can be uniquely expressed (up to the order of the factors) as a product of disjoint cycles, where each cycle is of length at least 2.*

Proof. We prove the result by induction on n.

Basis step: Suppose $n = 2$. Now $|S_2| = 2$ and the nonidentity element of S_2 is $\alpha = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$. Now $\alpha = (1\ 2)$, i.e., α is a cycle. Thus, the theorem is true for $n = 2$.

Inductive hypothesis: Suppose that the theorem is true for all S_k such that $2 \leq k < n$.

Inductive step: Suppose $n > 2$. We show that the result is true for n.

Let π be a nonidentity element of S_n. Now $\pi^i(1) \in I_n$ for all integers i, $i \geq 1$. Therefore, $\{\pi(1), \pi^2(1), \ldots, \pi^i(1), \ldots\} \subseteq I_n$. Because I_n is a finite set, we must have $\pi^l(1) = \pi^m(1)$ for some integers l and m such that $l > m \geq 1$. This implies that $\pi^{l-m}(1) = 1$. Let us write $j = l - m$. Then $j > 0$ and $\pi^j(1) = 1$. Let i be the smallest positive integer such that $\pi^i(1) = 1$. Let

$$A = \{1, \pi(1), \pi^2(1), \ldots, \pi^{i-1}(1)\}.$$

Then all elements of the set A are distinct. Let $\tau \in S_n$ be the permutation defined by

$$\tau = (1\ \pi(1)\ \pi^2(1)\ \cdots\ \pi^{i-1}(1)),$$

i.e., τ is a cycle. Let $B = I_n \backslash A$. If $B = \emptyset$, then π is a cycle. Suppose $B \neq \emptyset$. Let $\sigma = \pi|_B$. If σ is the identity, then π is a cycle. Suppose that σ is not the identity. Now by the induction hypothesis, σ is a product of disjoint cycles on B, say, $\sigma = \sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_r$. Now for $1 \leq i \leq r$, define π_i by

$$\pi_i(a) = \begin{cases} \sigma_i(a) \text{ if } a \in B \\ a \text{ if } a \notin B. \end{cases}$$

Then $\pi_1, \pi_2, \ldots, \pi_r$ and τ are disjoint cycles in S_n. It is easy to see that $\pi = \pi_1 \circ \pi_2 \circ \cdots \circ \pi_r \circ \tau$. Thus, π is a product of disjoint cycles.

To prove the uniqueness, let $\pi = \pi_1 \circ \pi_2 \circ \cdots \circ \pi_r = \mu_1 \circ \mu_2 \circ \cdots \circ \mu_s$, a product of r disjoint cycles and also a product of s disjoint cycles, respectively. We show that every π_i is equal to some μ_j and every μ_k is equal to some π_t. Consider π_i, $1 \leq i \leq r$. Suppose $\pi_i = (i_1 i_2 \ldots i_l)$. Then $\pi(i_1) \neq i_1$. This implies that i_1 is moved by some μ_l. By the disjointness of the cycles, there exists unique μ_j, $1 \leq j \leq s$, such that i_1 appears as an element in μ_j. By reordering, if necessary, we may write $\mu_j = (i_1\ c_2\ \ldots\ c_m)$. Now

$$\begin{array}{ccccccccccc}
i_2 & = & \pi_i(i_1) & = & \pi(i_1) & = & \mu_j(i_1) & = & c_2 & & \\
i_3 & = & \pi_i(i_2) & = & \pi(i_2) & = & \pi(c_2) & = & \mu_j(c_2) & = & c_3 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
i_l & = & \pi_i(i_{l-1}) & = & \pi(i_{l-1}) & = & \pi(c_{l-1}) & = & \mu_j(c_{l-1}) & = & c_l.
\end{array}$$

If $l < m$, then $i_1 = \pi_i(i_l) = \pi(i_l) = \pi(c_l) = \mu_j(c_l) = c_{l+1}$, a contradiction. Thus, $l = m$. Hence, $\pi_i = \mu_j$ for some j, $1 \leq j \leq s$. Similarly, every $\mu_k = \pi_t$ for some t, $1 \leq t \leq r$. ■

Corollary 3.1.19 *Let $n \geq 2$. Any permutation π of S_n can be expressed as a product of transpositions.*

Proof. In view of the preceding theorem, it suffices to show that every k-cycle can be expressed as a product of transpositions. This fact is immediate from the following equations:

$$e = (1) = (1\ 2) \circ (1\ 2)$$

and for $k \geq 2$

$$(i_1\ i_2\ \cdots\ i_k) = (i_1\ i_k) \circ (i_1\ i_{k-1}) \circ \cdots \circ (i_1\ i_2),$$

where $\{i_1, i_2 \ldots, i_k\} \subseteq I_n$. ■

Let $\pi \in S_n$. Because S_n is a finite group, we know that $\circ(\pi)$ is finite. Thus, in order to find the order of π, we need to compute π, π^2, $\pi^3, \ldots$, until we find the first positive integer k such that $\pi^k = e$. Finding such a positive integer could be a tedious task. However, we can effectively make use of the decomposition of π as a product of disjoint cycles, compute the order of each cycle, which is nothing but the length of the cycle (Exercise 17, page 69) and from the order of the cycles deduce the order of π. We ask the reader to consider this problem in Exercise 18 (page 69).

Theorem 3.1.18 tells us that any permutation $\alpha \in S_n$, $n \geq 2$, can be written as a product of disjoint cycles. However, the theorem does not tell us how to find the disjoint cycles in the decomposition of α. Next, we illustrate how to find these cycles.

Let π be a permutation on I_n, $n \geq 2$. In order to express π as a product of disjoint cycles, first consider 1, $\pi(1)$, $\pi^2(1)$, $\pi^3(1)$, ... and find the smallest positive integer r such that $\pi^r(1) = 1$. Let

$$\sigma_1 = (1\ \pi(1)\ \pi^2(1)\ \cdots\ \pi^{r-1}(1)).$$

Then σ_1 is a cycle of length r. Let i be the first element of I_n not appearing in σ_1. Now consider i, $\pi(i)$, $\pi^2(i)$, $\pi^3(i)$, ... and find the smallest positive integer s such that $\pi^{s-1}(i) = i$. Let

$$\sigma_2 = (i\ \pi(i)\ \pi^2(i)\ \cdots\ \pi^{s-1}(i)).$$

Then σ_2 is a cycle of length s. Now

$$\{1, \pi(1), \pi^2(1), \ldots, \pi^{r-1}(1)\} \cap \{i, \pi(i), \pi^2(i), \ldots, \pi^{s-1}(i)\} = \emptyset,$$

for if $j \in \{1,\ \pi(1),\ \pi^2(1),\ \ldots,\ \pi^{r-1}(1)\} \cap \{i,\ \pi(i),\ \pi^2(i),\ \ldots,\ \pi^{s-1}(i)\}$, then $j = \pi^p(i)$ for some p, $1 \leq p < r$, and $j = \pi^k(1)$ for some k, $1 \leq k < s$. Thus, $\{1,\ \pi(1),\ \pi^2(1),\ \ldots,\ \pi^{r-1}(1)\} = \{i,\ \pi(i),\ \pi^2(i),\ \ldots,\ \pi^{s-1}(i)\}$, which is a contradiction. Hence, σ_1 and σ_2 are disjoint cycles. If $\{1,\ \pi(1),\ \pi^2(1),\ \ldots,\ \pi^{r-1}(1)\} \cup \{i,\ \pi(i),\ \pi^2(i),\ \ldots,\ \pi^{s-1}(i)\} \neq I_n$, then consider the first element of I_n not appearing in $\{1,\ \pi(1),\ \pi^2(1),\ \ldots,\ \pi^{r-1}(1)\} \cup \{i,\ \pi(i),\ \pi^2(i),\ \ldots,\ \pi^{s-1}(i)\}$ and continue the above process to construct the cycle σ_3. Because I_n is finite, the above process must stop with some cycle σ_m. Then $\pi = \sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_m$.

We illustrate the above procedure with the help of the following example.

Example 3.1.20 *Consider the permutation*

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

on I_7. Here $\pi(1) = 6$, $\pi^2(1) = \pi(6) = 7$, and $\pi^3(1) = \pi(7) = 1$. That is, $1 \xrightarrow{\pi} 6 \xrightarrow{\pi} 7 \xrightarrow{\pi} 1$. Hence, $\sigma_1 = (1\ 6\ 7)$ is a 3-cycle. Now 2 is the first element of I_7 not appearing in $(1\ 6\ 7)$. Also, $\pi(2) = 3$, $\pi^2(2) = \pi(3) = 5$, $\pi^3(2) = \pi(5) = 4$, and $\pi^4(2) = \pi(4) = 2$. That is, $2 \xrightarrow{\pi} 3 \xrightarrow{\pi} 5 \xrightarrow{\pi} 4 \xrightarrow{\pi} 2$. Hence, $\sigma_2 = (2\ 3\ 5\ 4)$ is a cycle of length 4. Now σ_1 and σ_2 are disjoint and $\pi = \sigma_1 \circ \sigma_2$.

While writing a permutation as a product of disjoint cycles, it is customary not to write cycles of length one in the product. Thus, if some element of I_n does not appear in any of the cycles, then it is assumed to be fixed. For example, if $\pi = (1\ 2\ 5) \circ (4\ 6) \in S_7$, then because 3 and 7 neither appear in $(1\ 2\ 5)$ nor in $(4\ 6)$, they are fixed, i.e., $\pi(3) = 3$ and $\pi(7) = 7$.

Given a permutation $\pi \in S_n$, $n \geq 2$, we can write π as a product of disjoint cycles. We can also write π as a product of transpositions. However, the representation of π as a product of transposition need

not be unique. For example, (1 2 3) = (1 3) ∘ (1 2) = (2 1) ∘ (2 3). Also, (1 3) = (1 2) ∘ (1 3) ∘ (2 3). That is, (1 3) can be written as a product of one transposition or as a product of three transpositions. However, we will show that the number of transpositions in any representation of a permutation is either even or odd, but not both. We now proceed to prove this result.

Consider the formal product

$$\begin{aligned}\mathcal{X} \;=\; \textstyle\prod_{1\le i<j\le n}(a_i-a_j) \;=\; & (a_1-a_2)(a_1-a_3)\cdots(a_1-a_n)\\ & (a_2-a_3)\cdots(a_2-a_n)\\ & \vdots\\ & (a_{n-1}-a_n).\end{aligned}$$

If $n = 4$, then $\mathcal{X} = (a_1-a_2)(a_1-a_3)(a_1-a_4)(a_2-a_3)(a_2-a_4)(a_3-a_4)$.

For any permutation $\pi \in S_n$, let

$$\pi(\mathcal{X}) = \prod_{1\le i<j\le n}(a_{\pi(i)}-a_{\pi(j)}).$$

Let us first examine $\sigma(\mathcal{X})$ for any transposition $\sigma \in S_n$.

Lemma 3.1.21 *Let $n \ge 2$. Let $\sigma = (i\ j) \in S_n$, $i < j$, be a transposition. Then $\sigma(\mathcal{X}) = -\mathcal{X}$.*

Proof. First consider the factor $(a_i - a_j)$ in the product $\mathcal{X}$. The corresponding factor in $\sigma(\mathcal{X})$ is $a_{\sigma(i)} - a_{\sigma(j)}$. Now

$$a_{\sigma(i)} - a_{\sigma(j)} = a_j - a_i = -(a_i - a_j).$$

Next, consider the factor $a_k - a_l$, where both k and l are neither equal to i nor equal to j. The corresponding factor in $\sigma(\mathcal{X})$ is $a_{\sigma(k)} - a_{\sigma(l)}$ and

$$a_{\sigma(k)} - a_{\sigma(l)} = a_k - a_l.$$

Thus, the factor $a_k - a_l$ remains unaltered. Now consider the factor $a_k - a_l$, where either k or l (but not both) is equal to i or j. Let $1 \le t \le n$. Suppose $t < i < j$. We have the pair of factors $(a_t - a_i)$ and $(a_t - a_j)$ in the product $\mathcal{X}$. The corresponding factors in $\sigma(\mathcal{X})$ are $a_{\sigma(t)} - a_{\sigma(i)}$ and $a_{\sigma(t)} - a_{\sigma(j)}$ and

$$(a_{\sigma(t)} - a_{\sigma(i)})(a_{\sigma(t)} - a_{\sigma(j)}) = (a_t - a_j)(a_t - a_i) = (a_t - a_i)(a_t - a_j).$$

Therefore, the product $(a_t - a_i)(a_t - a_j)$ remains unchanged. Now suppose $i < t < j$. Then we have the pair of factors $(a_i - a_t)$ and $(a_t - a_j)$ in the product $\mathcal{X}$. The corresponding factors in $\sigma(\mathcal{X})$ are $a_{\sigma(i)} - a_{\sigma(t)}$ and $a_{\sigma(t)} - a_{\sigma(j)}$ and

$$(a_{\sigma(i)} - a_{\sigma(t)})(a_{\sigma(t)} - a_{\sigma(j)}) = (a_j - a_t)(a_t - a_i) = (a_i - a_t)(a_t - a_j).$$

Hence, the product $(a_i - a_t)(a_t - a_j)$ remains unaltered. Finally, let $i < j < t$. Then we have the pair of factors $(a_i - a_t)$ and $(a_j - a_t)$ in the product $\mathcal{X}$. The corresponding factors in $\sigma(\mathcal{X})$ are $a_{\sigma(i)} - a_{\sigma(t)}$ and $a_{\sigma(j)} - a_{\sigma(t)}$ and

$$(a_{\sigma(i)} - a_{\sigma(t)})(a_{\sigma(j)} - a_{\sigma(t)}) = (a_j - a_t)(a_i - a_t) = (a_i - a_t)(a_j - a_t).$$

Therefore, the product $(a_i - a_t)(a_j - a_t)$ remains unaltered. Thus, all factors other than $a_i - a_j$ and $a_k - a_l$, where both k and l are neither equal to i nor equal to j, can be paired so that the product of factors under σ remains unaltered. Hence, it now follows that $\sigma(\mathcal{X}) = -\mathcal{X}$. ■

Theorem 3.1.22 *Let $n \ge 2$. Let $\pi \in S_n$. Suppose*

$$\pi = \sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_r = \tau_1 \circ \tau_2 \circ \cdots \circ \tau_s,$$

where σ_i, $\tau_j \in S_n$ are transpositions, $i = 1, 2, \ldots, r$, and $j = 1, 2, \ldots, s$. Then both r and s are either even or odd.

Proof. By Lemma 3.1.21, $\sigma_i(\mathcal{X}) = -\mathcal{X}$ and $\tau_j(\mathcal{X}) = -\mathcal{X}$ for all $i = 1, 2, \ldots, r$, and $j = 1, 2, \ldots, s$. First we compute $(\sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_r)(\mathcal{X})$. Now

$$\begin{aligned}(\sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_r)(\mathcal{X}) &= \sigma_1(\sigma_2(\cdots(\sigma_r(\mathcal{X})))) \\ &= (-1)^r\mathcal{X}.\end{aligned}$$

Similarly, $(\tau_1 \circ \tau_2 \circ \cdots \circ \tau_s)(\mathcal{X}) = (-1)^s\mathcal{X}$. Hence, $(-1)^r = (-1)^s$. Thus, both r and s are either even of odd. ■

By the above theorem, if $\pi \in S_n$, then π can be written as a product of either an even or an odd number of transpositions, but not both. This leads us to the following definition.

Definition 3.1.23 *Let $\pi \in S_n$. If π is a product of an even number of transpositions, then π is called an **even permutation;** otherwise π is called an **odd permutation**.*

Corollary 3.1.24 *Let $\pi \in S_n$ be a k-cycle. Then π is an even permutation if and only if k is odd.*

Proof. Let $\pi = (1\ 2\ \cdots\ k)$. Then $\pi = (1\ k) \circ (1\ k-1) \circ \cdots \circ (1\ 2)$, i.e., π is a product of $k-1$ transposition. If π is an even permutation then $k-1$ is even, so k is odd. On the other hand, if k is odd, then $k-1$ is even, so π is an even permutation. This completes the proof. ■

Let A_n denote the subset of S_n consisting of all even permutations, $n \geq 2$.

Theorem 3.1.25 *For $n \geq 2$, the pair $(A_n, \circ)$ is a group, called the **alternating group on** I_n.*

Proof. Because $e = (1\ 2) \circ (1\ 2)$, $e \in A_n$. Thus, $A_n \neq \emptyset$. A product $\pi_1 \circ \pi_2$ is even if and only if π_1 and π_2 are both even or both odd by Theorem 3.1.22. Therefore, A_n is closed under $\circ$. If $\pi \in A_n$, then $\pi \circ \pi^{-1} = e$ is even and hence $\pi^{-1} \in A_n$. Hence, $(A_n, \circ)$ is a group. ■

Cauchy recognized many important properties of A_n. Among others, he proved the following theorem.

Theorem 3.1.26 *Every element in A_n is a product of 3-cycles, $n \geq 3$.*

Proof. Let $\pi \in A_n$. Then $\pi = \sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_r$, where σ_i is a transposition, $1 \leq i \leq r$, and r is even. Now for any transposition $(a\ b)$,

$$(a\ b) = (1\ a) \circ (1\ b) \circ (1\ a).$$

Thus,

$$\pi = (1\ i_1) \circ (1\ i_2) \circ \cdots \circ (1\ i_m)$$

where m is even. Because $(1\ i_1) \circ (1\ i_2) = (1\ i_2\ i_1)$, it follows that π is a product of 3-cycles. ■

Worked-Out Exercises

♢ **Exercise 1** Prove that two cycles in S_n are conjugate if and only if they have the same length.

Solution: Let $\alpha = (i_1 i_2 \cdots i_r)$ and $\beta = (j_1 j_2 \cdots j_s)$ be two cycles in S_n. First suppose that α and β are conjugate. Then $\beta = \sigma^{-1} \circ \alpha \circ \sigma$ for some $\sigma \in S_n$. Because σ is onto and $i_l \in I_n$, there exists k_l such that $\sigma(k_l) = i_l$ for all $l = 1, 2, \ldots, r$. Now

$$\begin{aligned}(j_1 j_2 \cdots j_s) &= (\sigma^{-1}(i_1)\sigma^{-1}(i_2) \cdots \sigma^{-1}(i_r)) \qquad (\text{ by Theorem 3.1.15}) \\ &= (k_1 k_2 \cdots k_r).\end{aligned}$$

Hence, $s = r$, so α and β are of the same length.

Conversely, let $\alpha = (i_1 i_2 \cdots i_r)$ and $\beta = (j_1 j_2 \cdots j_r)$ be two cycles in S_n of the same length. Let $\sigma = \begin{pmatrix} i_1 & i_2 & \cdots & i_r \\ j_1 & j_2 & \cdots & j_r \end{pmatrix}$, i.e., $\sigma(i_l) = j_l$ for all $l = 1, 2, \ldots, r$, and $\sigma(a) = a$ for all $a \in I_n \backslash \{i_1, i_2, \ldots, i_r\}$. Then $\sigma \in S_n$. Now

$$\sigma^{-1} \circ \beta \circ \sigma = (\sigma^{-1}(j_1)\sigma^{-1}(j_2) \cdots \sigma^{-1}(j_r)) = (i_1 i_2 \cdots i_r) = \alpha.$$

♢ **Exercise 2** Express the permutation

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

on I_8 as a product of disjoint cycles and then as a product of transposition. Is σ an even permutation?

Solution: We have $\sigma(1) = 2$, $\sigma^2(1) = \sigma(2) = 3$, $\sigma^3(1) = \sigma(3) = 8$, and $\sigma^4(1) = \sigma(8) = 1$. Thus, (1 2 3 8) is a cycle. Now 4 is the first element of I_8 not appearing in (1 2 3 8). We have $\sigma(4) = 5$, $\sigma^2(4) = \sigma(5) = 6$, and $\sigma^3(4) = \sigma(6) = 4$. Hence, (4 5 6) is also a cycle in σ. Next, 7 is the first element of I_8 not appearing in (1 2 3 8) and (4 5 6). Now $\sigma(7) = 7$. Because all the elements of I_8 appear in one of the cycles (1 2 3 8), (4 5 6), and (7), we have $\sigma = (1\ 2\ 3\ 8) \circ (4\ 5\ 6)$. Now (1 2 3 8) = (1 8) ∘ (1 3) ∘ (1 2) and (4 5 6) = (4 6) ∘ (4 5). Thus, $\sigma = (1\ 8) \circ (1\ 3) \circ (1\ 2) \circ (4\ 6) \circ (4\ 5)$. Because σ is a product of five transpositions, σ is not an even permutation.

♢ **Exercise 3** Write all elements of S_4. Show that S_4 has no elements of order ≥ 5.

Solution: Let $\sigma \in S_4$ and $\sigma = \sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_k$, a product of disjoint cycles. Because S_4 is a permutation group on I_4, $k \leq 2$. If $k = 1$, then σ is a 2-cycle, 3-cycle, or 4-cycle. If $k = 2$, then σ is a product of two disjoint transpositions. The number of distinct cycles of length 2 is 6, the number of distinct cycles of length 3 is 8, and the number of distinct cycles of length 4 is 6. Hence, $S_4 = \{e$, (1 2), (1 3), (1 4), (2 3), (2 4), (3 4), (1 2 3), (1 3 2), (2 3 4), (2 4 3), (1 3 4), (1 4 3), (1 2 4), (1 4 2), (1 2 3 4), (1 3 2 4), (1 4 2 3), (1 2 4 3), (1 3 4 2), (1 4 3 2), (1 2) ∘ (3 4), (1 4) ∘ (3 2), (1 3) ∘ (2 4)$\}$.

Because each 2-cycle is of order 2, each 3-cycle is of order 3, each 4-cycle is of order 4, and the order of the product of two disjoint 2-cycles is 2, S_4 has no element of order ≥ 5.

♢ **Exercise 4** Find the order of (1 2 3 4) ∘ (5 6 7) in S_7.

Solution: ∘(1 2 3 4) = 4, ∘(5 6 7) = 3. Now (1 2 3 4) and (5 6 7) are disjoint. Hence, (1 2 3 4) ∘ (5 6 7) = (5 6 7) ∘ (1 2 3 4). If a and b are two elements of a group G such that ∘(a) = m, ∘(b) = n, and $\gcd(m, n) = 1$, then ∘(ab) = mn. Using this result, we find that the order of (1 2 3 4) ∘ (5 6 7) is 12.

♢ **Exercise 5** Find the order of (1 2 3 4) ∘ (5 6) in S_6.

Solution: ∘(1 2 3 4) = 4, ∘(5 6) = 2. Now (1 2 3 4) and (5 6) are disjoint, so they commute. Thus, ((1 2 3 4) ∘ (5 6))$^4 = e$. Now ((1 2 3 4) ∘ (5 6))$^1 \neq e$, ((1 2 3 4) ∘ (5 6))^2 = (1 2 3 4)^2 ∘ (5 6)^2 = (1 2 3 4)$^2 \neq e$. If ((1 2 3 4) ∘ (5 6))$^3 = e$, then the order of (1 2 3 4) ∘ (5 6) will be 3 and 3 divides 4, a contradiction. Hence, the order of (1 2 3 4) ∘ (5 6) is 4.

Exercises

1. Express the following permutations as (i) a product of disjoint cycles and (ii) a product of transpositions:
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 5 & 4 & 1 & 6 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 1 & 5 & 4 & 6 \end{pmatrix}.$$
2. Let $\alpha = (1\ 2\ 5\ 7)$ and $\beta = (2\ 4\ 6) \in S_7$. Find $\alpha \circ \beta \circ \alpha^{-1}$.
3. Let $\alpha = (1\ 3\ 5\ 7)$ and $\beta = (2\ 4\ 8) \circ (1\ 3\ 6) \in S_8$. Find $\alpha \circ \beta \circ \alpha^{-1}$.
4. Let $\alpha = (1\ 3) \circ (5\ 8)$ and $\beta = (2\ 3\ 6\ 7) \in S_8$. Find $\alpha \circ \beta \circ \alpha^{-1}$.
5. Let $\alpha = (2\ 5\ 9) \circ (1\ 3\ 6)$ and $\beta = (1\ 5\ 7) \circ (2\ 4\ 6\ 9) \in S_9$. Find $\alpha \circ \beta \circ \alpha^{-1}$.
6. Let (1 3 5 7) and (2 3 6 8) $\in S_8$. Find $\alpha \in S_8$ such that $\alpha \circ (1\ 3\ 5\ 7) \circ \alpha^{-1} = (2\ 3\ 6\ 8)$.
7. If $\alpha = (1\ 2\ 3\ 4\ 5\ 6)$, show that $\alpha = (1\ 6) \circ (1\ 5) \circ (1\ 4) \circ (1\ 3) \circ (1\ 2)$.

8. Find the order of $(1\ 2\ 3) \circ (4\ 5)$ in S_5.

9. Prove that $(1\ 2 \cdots\ n-1\ n)^{-1} = (n\ n-1 \cdots 2\ 1)$.

10. Prove that every transposition is its own inverse.

11. Prove that the symmetric group on two symbols $(S_2, \circ)$ is commutative.

12. Let $\alpha = (a_1\ a_2\ \cdots\ a_k) \in S_n$ be a k-cycle. Show that

$$\alpha^2 = \begin{cases} (a_1\ a_3\ \cdots\ a_{2m-1}) \circ (a_2\ a_4\ a_6\ \cdots\ a_{2m}) \text{ if } k = 2m, \text{ i.e., } k \text{ is even} \\ (a_1\ a_3\ \cdots\ a_{2m+1}\ a_2\ a_4\ \cdots a_{2m}) \text{ if } k = 2m+1, \text{ i.e., } k \text{ is odd.} \end{cases}$$

13. Determine A_4.

14. Let $\alpha, \beta \in S_n$. Show that $\alpha^{-1} \circ \beta^{-1} \circ \alpha \circ \beta \in A_n$.

15. Prove that $|A_n| = \frac{n!}{2}$.

16. Show that the number of distinct cycles of length r in S_n is $\frac{1}{r}\frac{n!}{(n-r)!}$.

17. Let $n \geq 2$ and $\sigma \in S_n$ be a cycle. Show that σ is a k-cycle if and only if $\circ(\sigma) = k$.

18. Let $\sigma \in S_n$ and $\sigma = \sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_k$ be a product of disjoint cycles. Suppose $\circ(\sigma_i) = n_i$, $i = 1, 2, \ldots, n$. Show that $\circ(\sigma) = \text{lcm}(n_1, n_2, \ldots, n_k)$.

19. Let $\alpha \in S_n$ and p be a prime.

    (a) Show that $\circ(\alpha) = p$ if and only if either α is a p-cycle or α is a product of disjoint cycles, where each cycle is either of length 1 or length p and at least one cycle is of length p.

    (b) If α is a p-cycle, prove that either $\alpha^m = e$ or α^m is a p-cycle for all $m \in \mathbb{N}$.

20. Let α and $\beta \in S_n$. Let $\alpha = \alpha_1 \circ \alpha_2 \circ \cdots \circ \alpha_k$ and $\beta = \beta_1 \circ \beta_2 \circ \cdots \circ \beta_s$ be a product of disjoint cycles. Let $\text{length}(\alpha_i) = d_i$ and $\text{length}(\beta_j) = m_j$ for all $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, s$ and $d_1 \leq d_2 \leq \cdots \leq d_k$ and $m_1 \leq m_2 \leq \cdots \leq m_s$. We say that α and β have the same **cyclic structure** if $k = s$ and $d_i = m_i$ for all $i = 1, 2, \ldots, k$. Prove that α and β have the same cyclic structure if and only if α and β are conjugate.

21. Prove that for $\pi \in S_n$, π is an even permutation if and only if $\pi(\mathcal{X}) = \mathcal{X}$.

22 (a) Let $\alpha = (k\ l)$, $\beta \in S_n$ be two distinct transpositions, $n \geq 3$. Show that there exist transpositions $\mu, \nu \in S_n$ such that $\beta \circ \alpha = \nu \circ \mu$, $\mu(k) = k$ and ν moves k.

    (b) Prove that if the identity permutation $e \in S_n$ can be written as a product of r (≥ 3) transpositions, then e can be written as a product of $r-2$ transpositions.

    (c) Prove that if $e = \sigma_1 \circ \sigma_2 \circ \cdots \circ \sigma_r \in S_n$ as a product of transpositions, then r is even.

    (d) Use (a), (b), and (c) to prove that if $\pi \in S_n$, then π can be written as a product of either an even or an odd number of transpositions, but not both.

Augustin-Louis Cauchy (1789– 1857) was born on August 21, 1789, in Paris, France. He received his first education from his father. He was a neighbor of Laplace and Berthollet. Cauchy became acquainted with famous scientists at a young age. Lagrange is said to have warned his father not to show Cauchy any mathematics book before the age of seventeen.

At the age of fifteen, he completed his classic studies with distinction. He became an engineer in 1810, in the Napoleon army. In 1813, he returned to Paris.

In 1811, Cauchy started his mathematical career by solving a problem sent to him by Lagrange on convex polygons. In 1812, he solved Fermat's famous classical problem on polygon numbers. His treatise on the definite integral, which he submitted in 1814 to the French Academy, later became a basis of the theory of complex functions.

In 1816, he was appointed full professor at the École Polytechnique. More theorems and concepts have been named for Cauchy than for any other mathematician. There are sixteen concepts and theorems named for Cauchy in elasticity alone.

He worked on mathematics, mathematical physics, and celestial mechanics. In mathematics, he worked on several areas, such as calculus, complex functions, algebra, differential equations, geometry, and analysis. The notion of continuity used today was invented by Cauchy. He also proved that a continuous function has a zero between two points where the function changes its signs, a result also proved by Bolzano. The first adequate definitions of indefinite integral and definite improper integral are due to Cauchy

In algebra, the notion of the order of an element, a subgroup, and conjugates are found in his papers. He proved the famous Cauchy's theorem for finite groups, that is, if the order of a finite group is divisible by a prime p, then the group has a subgroup of order p. Cauchy's role in shaping the theory of permutation groups is central. He is regarded by some to be the founder of finite group theory. The two-row notation for permutations was introduced by Cauchy. He also defined the product of permutations, inverse permutations, transpositions, and the cyclic notation. He wrote his first paper on this subject in 1815, but did not return to it for nearly thirty years. In 1844, he proved that every permutation is a product of disjoint cycles.

He also did work of fundamental importance in the theory of determinants. His treatise on determinants, published in 1812, contains important results concerning product theorems and the inverse of a matrix.

Cauchy enjoyed teaching. He published more than 800 papers and eight books. He died on May 22, 1857.

Chapter 4

Subgroups and Normal Subgroups

In Chapter 2, we began a discussion of the evolution of group theory. This chapter seems a good place to renew the discussion. It took more than 100 years for the abstract concept of a group to evolve. The evolution followed lines similar to the evolution of other theories. First came the discovery of isolated phenomena, followed by the recognition of features common to all. Then came the search and classification of other instances. Next, general principles emerged. Last, the abstract postulates which define the system were uncovered. A deeper account can be found in Bell.

4.1 Subgroups

Let us consider the groups $(\mathbb{Z}, +)$ and $(\mathbb{Q}, +)$, where $+$ is the usual addition of numbers, and note the following:

1. Both these groups have the same binary operation.
2. $\mathbb{Z}$ is a subset of $\mathbb{Q}$.

The same is true for the groups $(\mathbb{Z}, +)$ and $(\mathbb{R}, +)$; $(\mathbb{Q}, +)$ and $(\mathbb{R}, +)$; $(\mathbb{R}, +)$ and $(\mathbb{C}, +)$. Similarly, as seen in the previous chapter, both the groups $(A_n, \circ)$ and $(S_n, \circ)$ have the same binary operations and A_n is a subset of S_n.

One can think of many examples, where the underlying set of one group is a subset of the underlying set of another group. This leads us to the concept of a subgroup. Before formally defining subgroups, let us also note the following:

Let $(G, *)$ be a group and H be a nonempty subset of G. Then H is said to be *closed under the binary operation* $*$ if $a * b \in H$ for all $a, b \in H$.

Suppose H is closed under the binary operation $*$. Then the restriction of $*$ to $H \times H$ is a mapping from $H \times H$ into H. Thus, the binary operation $*$ defined on G induces a binary operation on H. We denote this induced binary operation on H by $*$ also. Thus, $(H, *)$ is a mathematical system. It also follows that $*$ is associative as a binary operation on H, i.e., $a * (b * c) = (a * b) * c$ for all $a, b, c \in H$. If $(H, *)$ is a group, then we call H a subgroup of G. More formally, we have the following definition.

Definition 4.1.1 *Let $(G, *)$ be a group and H be a nonempty subset of G. Then $(H, *)$ is called a **subgroup** of $(G, *)$ if $(H, *)$ is a group.*

Consider the group $(\mathbb{Q}, +)$ and its subgroups $(\mathbb{Z}, +)$. Now the identity elements of both these groups is 0. Next, let $a \in \mathbb{Z}$. Then $a \in \mathbb{Q}$. Also, the inverse of a in $\mathbb{Z}$ as well as in $\mathbb{Q}$ is $-a$. In other words, the inverse of a in $\mathbb{Z}$ and the inverse of a in $\mathbb{Q}$ is the same. In general, we have the following result.

Theorem 4.1.2 *Let $(G, *)$ be a group and $(H, *)$ be a subgroup of $(G, *)$.*
(i) The identity elements of $(H, *)$ and $(G, *)$ are the same.
(ii) If $h \in H$, then the inverse of h in H and the inverse of h in G is the same.

Proof. (i) Let e_H denote the identity of H and e denote the identity of G. Note that

$$e_H * e_H = e_H = e_H * e.$$

Hence, by the cancellation property, $e_H = e$. This implies that the identity elements of G and H are the same.

(ii) Let $h \in H$. Let h' denote the inverse of h in H and h^{-1} denote the inverse of h in G. Then

$$h * h' = e = h' * h$$

and

$$h * h^{-1} = e = h^{-1} * h.$$

Now

$$h' = h' * e = h' * (h * h^{-1}) = (h' * h) * h^{-1} = e * h^{-1} = h^{-1}.$$

This implies that the inverse of h in H and the inverse of h in G are the same. ■

Remark 4.1.3 *If $(G, *)$ is a group, then $(\{e\}, *)$ and $(G, *)$ are subgroups of $(G, *)$. These subgroups are called* ***trivial****.*

Example 4.1.4 *Consider the following list of groups.*
(i) $(\{0\}, +)$, $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, $(\mathbb{C}, +)$,
(ii) $(\{1\}, \cdot)$, $(\mathbb{Q}\backslash\{0\}, \cdot)$, $(\mathbb{R}\backslash\{0\}, \cdot)$, $(\mathbb{C}\backslash\{0\}, \cdot)$,
where $+$ is the usual addition operation and $\cdot$ is the usual multiplication operation. Each group is a subgroup of the group listed to its right. For example, $(\mathbb{Z}, +)$ is a subgroup of $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$, and $(\mathbb{R}\backslash\{0\}, \cdot)$ is a subgroup of $(\mathbb{C}\backslash\{0\}, \cdot)$.

Notation 4.1.5 *In the remainder of the text, we shall generally use the notation G instead of $(G, *)$ for a group and we write ab for $a * b$. We shall refer to ab as the product of a and b. This notation is usually called* ***multiplicative notation****.*

Readers with some knowledge of linear algebra should notice the similarity with respect to the type of results and order of presentation of those which immediately follow. First comes a result which gives an easy method of determining if a nonempty subset is a substructure. This is followed by the result that the intersection of any collection of substructures is a substructure. Next, comes the definition of a substructure "generated" by a subset. Finally, a theorem describing the substructure generated by a given subset. These ideas appear throughout algebra. We will encounter them again, for example, when we examine the ideals of a ring.

Let G be a group and H be a nonempty subset of G. To show that H is a subgroup of G, we need to show that H is a group under the binary operation of G. This requires us to show that H is closed under the binary operation, the binary operation is associative, H contains the idenntiy element, and every element of H has an inverse in H. Next theorem gives a criteria that can be effectively used to show that a nonempty subset of a group is a subgroup.

Theorem 4.1.6 *Let G be a group and H be a nonempty subset of G. Then H is a subgroup of G if and only if for all $a, b \in H$, $ab^{-1} \in H$.*

Proof. Suppose H is a subgroup of G. Let a, $b \in H$. Because H is a subgroup, it is a group. Therefore, $b \in H$ implies that $b^{-1} \in H$. Thus, $ab^{-1} \in H$ because H is closed under the binary operation.

Conversely, suppose H is a nonempty subset of G such that a, $b \in H$ implies $ab^{-1} \in H$. Because $H \neq \emptyset$, there exists $a \in H$. Now $a, a \in H$. Therefore, $e = aa^{-1} \in H$, i.e., H contains the identity. Next, let $b \in H$. Then $e, b \in H$, implies that $b^{-1} = eb^{-1} \in H$. Thus, every element of H has an inverse in H.

To show that H is closed under the binary operation, let $a, b \in H$. Then a, $b^{-1} \in H$. Thus, $ab = a(b^{-1})^{-1} \in H$. Hence, H is closed under the binary operation. From the statements preceding Definition 4.1.1, associativity holds for H. Hence, H is a group, so H is subgroup of G.
■

Corollary 4.1.7 *Let G be a group and H be a finite nonempty subset of G. Then H is a subgroup of G if and only if for all $a, b \in H$, $ab \in H$.*

Proof. If H is a subgroup, then for all $a, b \in H$, $ab \in H$.

Conversely, suppose that for all $a, b \in H$, $ab \in H$. Let $h \in H$. Then $h, h^2, \ldots, h^n, \ldots \in H$, so

$$\{h, h^2, \ldots, h^n, \ldots\} \subseteq H.$$

Because H is finite and the set $\{h, h^2, \ldots, h^n, \ldots\}$ is a subset of H, it follows that all the elements of $\{h, h^2, \ldots, h^n, \ldots\}$ cannot be distinct. Thus, there exist integers r and s such that $0 \leq r < s$ and $h^r = h^s$. This implies that $e = h^{s-r} \in H$. Now $s - r \geq 1$. Thus, $e = hh^{s-r-1}$ implies that $h^{-1} = h^{s-r-1} \in H$.

Let $a, b \in H$. Then $a, b^{-1} \in H$, so $ab^{-1} \in H$ by the hypothesis. Hence, by Theorem 4.1.6, H is a subgroup. ■

Theorem 4.1.8 *Let G be a group and $Z(G) = \{b \in G \mid ab = ba \text{ for all } a \in G\}$. Then $Z(G)$ is a commutative subgroup of G. $Z(G)$ is called the* ***center*** *of G.*

Proof. Since $ae = a = ea$ for all $a \in G$, $e \in Z(G)$ and so $Z(G) \neq \emptyset$. Let $a, b \in Z(G)$. Then $bc = cb$ for all $c \in G$. From this, it follows that $cb^{-1} = b^{-1}c$ for all $c \in G$ and so $b^{-1} \in Z(G)$. Now

$$(ab^{-1})c = a(b^{-1}c) = a(cb^{-1}) = (ac)b^{-1} = (ca)b^{-1} = c(ab^{-1})$$

for all $c \in G$ and so $ab^{-1} \in Z(G)$. Hence by Theorem 4.1.6, $Z(G)$ is a subgroup of G. That $Z(G)$ is commutative follows by the definition of $Z(G)$. ■

In the remainder of this section, we will see how new subgroups arise from existing subgroups of a group.

Theorem 4.1.9 *Let G be a group and $\{H_\alpha \mid \alpha \in I\}$ be any nonempty collection of subgroups of G. Then $\cap_{\alpha \in I} H_\alpha$ is a subgroup of G.*

Proof. Each H_α is a subgroup. Therefore, $e \in H_\alpha$ for all $\alpha \in I$. This implies that $e \in \cap_{\alpha \in I} H_\alpha$, so $\cap_{\alpha \in I} H_\alpha \neq \emptyset$. To show that $\cap_{\alpha \in I} H_\alpha$ is a subgroup, we will use Theorem 4.1.6.

Let $a, b \in \cap_{\alpha \in I} H_\alpha$. Now

$$\begin{array}{ll} & a, b \in \cap_{\alpha \in I} H_\alpha \\ \Rightarrow & a, b \in H_\alpha \text{ for all } \alpha \in I \\ \Rightarrow & ab^{-1} \in H_\alpha \text{ for all } \alpha \in I \quad \text{because } H_\alpha \text{ is a subgroup for all } \alpha \in I \\ \Rightarrow & ab^{-1} \in \cap_{\alpha \in I} H_\alpha. \end{array}$$

Consequently, $\cap_{\alpha \in I} H_\alpha$ is a subgroup of G by Theorem 4.1.6. ■

Definition 4.1.10 *Let G be a group and S be a subset of G. Let*

$$\mathcal{S} = \{H \mid H \text{ is a subgroup of } G \text{ and } S \subseteq H\}.$$

Define

$$\langle S \rangle = \cap_{H \in \mathcal{S}} H,$$

i.e., $\langle S \rangle$ is the intersection of all subgroups H of G such that $S \subseteq H$.

(i) Thesubgroup $\langle S \rangle$ of G is called the ***subgroup generated*** *by S.*

(ii) If $G = \langle S \rangle$, then S is called a set of ***generators*** *for G.*

Remark 4.1.11 *Let G be a group and S be a subset of G. Note that if either $S = \emptyset$ or $S = \{e\}$, then $\langle S \rangle = \{e\}$. Moreover, $\langle G \rangle = G$.*

We now proceed to obtain a characterization of a subgroup generated by a nonempty subset in terms of the elements of the group.

Let $\mathcal{S} = \{H \mid H \text{ is a subgroup of } G \text{ and } S \subseteq H\}$, where $S \neq \emptyset$. It can be shown that $(\mathcal{S}, \leq)$ is a partially ordered set, where $\leq$ denotes the set inclusion relation. In this poset, $\langle S \rangle$ is the least element. Hence, $\langle S \rangle$ is the smallest subgroup of G which contains S.

Because $\langle S \rangle$ is a subgroup of G, we must have for any $s_1, \ldots, s_n \in S$, the product $s_1^{k_1} \cdots s_n^{k_n} \in \langle S \rangle$, where $k_i = \pm 1$ for $i = 1, 2, \ldots, n$. Thus, if A denotes the set $\{s_1^{k_1} \cdots s_n^{k_n} \mid s_i \in S,\ k_i = \pm 1, i = 1, 2, \ldots, n;\ n = 1, 2, \ldots\}$, then $A \subseteq \langle S \rangle$. Note that if $s \in S$, then $e = ss^{-1} \in A$. In the following theorem, we show that $A = \langle S \rangle$. Therefore, S does "generate" $\langle S \rangle$ in the sense of multiplying elements of S or their inverses together to build up the smallest subgroup containing S.

Theorem 4.1.12 *Let S be a nonempty subset of a group G. Then*

$$\langle S \rangle = \{s_1^{k_1} \cdots s_n^{k_n} \mid s_i \in S,\ k_i = \pm 1,\ i = 1, 2, \ldots, n;\ n = 1, 2, \ldots\}.$$

Proof. Let

$$A = \{s_1^{k_1} \cdots s_n^{k_n} \mid s_i \in S,\ k_i = \pm 1, i = 1, 2, \ldots, n;\ n = 1, 2, \ldots\}.$$

We have already noted that $A \subseteq \langle S \rangle$. We show that $\langle S \rangle \subseteq A$ by showing that A is a subgroup of G containing S. (Recall that $\langle S \rangle$ is the smallest subgroup of G containing S.)

Because $S \neq \emptyset$, there exists $s \in S$. Then $s = s^1 \in A$, so $S \subseteq A$. Let $a = s_1^{i_1} \cdots s_m^{i_m}$, $b = t_1^{j_1} \cdots t_q^{j_q} \in A$. Then

$$ab^{-1} = (s_1^{i_1} \cdots s_m^{i_m})(t_1^{j_1} \cdots t_q^{j_q})^{-1} = s_1^{i_1} \cdots s_m^{i_m} t_q^{-j_q} \cdots t_1^{-j_1} \in A.$$

Thus, A is a subgroup of G by Theorem 4.1.6. Hence, $\langle S \rangle \subseteq A$.

■

Notation 4.1.13 *For $a \in G$, we use the notation $\langle a \rangle$ rather than $\langle \{a\} \rangle$ to denote the subgroup of G generated by $\{a\}$.*

Corollary 4.1.14 *Let G be a group and $a \in G$. Then $\langle a \rangle = \{a^n \mid n \in \mathbb{Z}\}$.*

Proof. By Theorem 4.1.12, we have $\langle a \rangle = \{a^{k_1} \cdots a^{k_m} \mid k_i = \pm 1,\ i = 1, 2, \ldots,\ m;\ m = 1, 2, \ldots\} = \{a^{k_1 + \cdots + k_m} \mid k_i = \pm 1,\ i = 1, 2, \ldots,\ m;\ m = 1, 2, \ldots\} = \{a^n \mid n \in \mathbb{Z}\}$.
■

In the additive notation, we would have $\langle a \rangle = \{na \mid n \in \mathbb{Z}\}$.

Let $n \geq 3$. From Chapter 3, recall that A_n is the set of all even permutations on the set I_n. Moreoevr, in Chapter 3, we proved that every element of A_n is a product of 3-cycles (Theorem 3.1.26). In the following theorem, we conclude that A_n is generated by the set of all 3-cycles.

Theorem 4.1.15 *Let $n \geq 3$. Then A_n is generated by the set of all 3-cycles.*

Proof. Let $(a\ b\ c)$ be a 3-cycle. Then $(a\ b\ c) = (a\ c) \circ (a\ b)$, i.e., $(a\ b\ c)$ is a product of even number of transposition. Hence, every 3-cycle is an even permutation. This implies that every 3-cycle is in A_n. By Theorem 3.1.26, every element of A_n is a product of 3-cycles. Hence, A_n is generated by the set of all 3-cycles.
■

We now turn our attention to the product of subgroups.

Definition 4.1.16 *Let H and K be nonempty subsets of a group G. The product of H and K is defined to be the set*

$$HK = \{hk \mid h \in H, k \in K\}.$$

Let $H_1, H_2, \ldots, H_n$ be nonempty subsets of a group G. We define the product, $H_1 H_2 \cdots H_n$, of $H_1, H_2, \ldots, H_n$ to be the set

$$H_1 H_2 \cdots H_n = \{h_1 h_2 \cdots h_n \mid h_i \in H_i,\ i = 1, 2, \ldots, n\}.$$

Example 4.1.17 *Consider the group of symmetries of the square. Let $H = \{r_{360}, d_1\}$ and $K = \{r_{360}, h\}$. Then H and K are subgroups of G. Now*

$$HK = \{r_{360}r_{360}, r_{360}h, d_1 r_{360}, d_1 h\} = \{r_{360}, h, d_1, r_{90}\}.$$

Now $hd_1 = r_{270} \notin HK$, so HK is not closed under the binary operation. Hence, HK is not a subgroup of the symmetries of the square. Also, note that

$$KH = \{r_{360}r_{360}, r_{360}d_1, hr_{360}, hd_1\} = \{r_{360}, d_1, h, r_{270}\},$$

and

$$\langle H \cup K \rangle = \{r_{360}, r_{90}, r_{180}, r_{270}, h, v, d_1, d_2\}.$$

Example 4.1.17 shows that in general the product of subgroups need not be a subgroup. In the following theorem, we give a necessary and sufficient condition for the product of subgroups to be a subgroup.

Theorem 4.1.18 *Let H and K be subgroups of a group G. Then HK is a subgroup of G if and only if $HK = KH$.*

Proof. Suppose HK is a subgroup of G. Let $kh \in KH$, where $h \in H$ and $k \in K$. Now $h = he \in HK$ and $k = ek \in HK$. Because HK is a subgroup, it follows that $kh \in HK$. Hence, $KH \subseteq HK$. On the other hand, let $hk \in HK$. Then $(hk)^{-1} \in HK$, so $(hk)^{-1} = h_1 k_1$ for some $h_1 \in H$ and $k_1 \in K$. Thus,

$$hk = (h_1 k_1)^{-1} = k_1^{-1} h_1^{-1} \in KH.$$

This implies that $HK \subseteq KH$. Hence, $HK = KH$.

Conversely, suppose $HK = KH$. Let $h_1 k_1,\ h_2 k_2 \in HK$, where $h_1,\ h_2 \in H$ and $k_1,\ k_2 \in K$. We show that $(h_1 k_1)(h_2 k_2)^{-1} \in HK$.

Now $k_2 \in K$ and $h_2 \in H$. Therefore, $k_2^{-1} h_2^{-1} \in KH = HK$. This implies that $k_2^{-1} h_2^{-1} = h_3 k_3$ for some $h_3 \in H$ and $k_3 \in K$. Similarly, $k_1 h_3 \in KH = HK$, so $k_1 h_3 = h_4 k_4$ for some $h_4 \in H$ and $k_4 \in K$. Thus,

$$\begin{array}{lcll} (h_1 k_1)(h_2 k_2)^{-1} & = & h_1 k_1 k_2^{-1} h_2^{-1} & \text{because } (h_2 k_2)^{-1} = k_2^{-1} h_2^{-1} \\ & = & h_1 k_1 h_3 k_3 & \text{substitute } k_2^{-1} h_2^{-1} = h_3 k_3 \\ & = & h_1 h_4 k_4 k_3 \in HK. & \text{substitute } k_1 h_3 = h_4 k_4 \end{array}$$

Hence, HK is a subgroup of G by Theorem 4.1.6. ■

Corollary 4.1.19 *If H and K are subgroups of a commutative group G, then HK is a subgroup of G.*

Proof. Since G is commutative, $HK = KH$. The result now follows by Theorem 4.1.18 ■

The following theorem gives another necessary and sufficient condition for a product of subgroups to be a subgroup.

Theorem 4.1.20 *Let H and K be subgroups of a group G. Then HK is a subgroup of G if and only if $HK = \langle H \cup K \rangle$.*

Proof. First suppose that HK is a subgroup of G. We show that $HK = \langle H \cup K \rangle$.

Let $h \in H$. Then $h = he \in HK$. Thus, $H \subseteq HK$. Similarly, $K \subseteq HK$. Hence, $H \cup K \subseteq HK$. Now $\langle H \cup K \rangle$ is the smallest subgroup of G containing $H \cup K$, so it follows that $\langle H \cup K \rangle \subseteq HK$.

Let $hk \in HK$, where $h \in H$ and $k \in K$. Because $H \subseteq \langle H \cup K \rangle$ and $K \subseteq \langle H \cup K \rangle$, we have $h, k \in \langle H \cup K \rangle$. Thus, $hk \in \langle H \cup K \rangle$. This implies that $HK \subseteq \langle H \cup K \rangle$. Hence, $HK = \langle H \cup K \rangle$.

The converse is immediate because $\langle H \cup K \rangle$ is a subgroup and $HK = \langle H \cup K \rangle$. ■

Example 4.1.21 *Let $G = \langle a, b \rangle$, where $a^3 = e$, $b^2 = e$, and $(ab)^2 = e$. Then*

(i) $ab = ba^{-1}$, $ba = a^{-1}b$, and $a^2b = ba$.

(ii) G is not commutative because $ab \neq ba$.

(iii) $ba^s = a^{-s}b$ for all positive integers s.

(iv) By (i) and (iii)

$$a^r b^i a^s b^j = \begin{cases} a^{r+s}b^j & \text{if } i = 0 \\ a^{r-s}b^{i+j} & \text{if } i = 1. \end{cases}$$

(v) Because $a^3 = e = b^2$, every element of G is of the form $a^r b^i$, $0 \leq r < 3$, $i = 0, 1$ by (iv).

(vi) $G = \{e, a, b, ab, a^2, a^2b\}$. Thus, $|G| = 6$.

(vii) $\circ(a) = 3 = \circ(a^2)$, $\circ(b) = \circ(ab) = \circ(a^2b) = 2$.

(viii) The only subgroups of G are $\{e\}$, $\langle a \rangle = \langle a^2 \rangle$, $\langle b \rangle$, $\langle ab \rangle$, $\langle a^2b \rangle$, and G.

*G is called a **dihedral group** of degree 3 and is denoted by D_3. In general, a dihedral group[1] of degree n is $D_n = \langle a, b \rangle$, where $(ab)^2 = e$, $\circ(a) = n$, and $\circ(b) = 2$. In Chapter 5, we consider, D_4, a dihedral group of degree 4, and study this group in detail.*

Worked-Out Exercises

♢ **Exercise 1** Let H be a subgroup of a group G. Let $g \in G$. Prove that

(a) $gHg^{-1} = \{ghg^{-1} \mid h \in H\}$ is a subgroup of G,

(b) $\left|gHg^{-1}\right| = |H|$.

Solution (a) We first show that $gHg^{-1} \neq \emptyset$ and then use Theorem 4.1.6. Since $e = geg^{-1} \in gHg^{-1}$, $gHg^{-1} \neq \emptyset$. Let $gh_1g^{-1}, gh_2g^{-1} \in gHg^{-1}$. Then

$$(gh_1g^{-1})(gh_2g^{-1})^{-1} = gh_1g^{-1}gh_2^{-1}g^{-1} = gh_1h_2^{-1}g^{-1} \in gHg^{-1}.$$

Hence, gHg^{-1} is a subgroup of G.

(b) Let $g \in G$. To prove that $\left|gHg^{-1}\right| = |H|$, we show that there exists a one-one onto function of H onto gHg^{-1}. Define $f : H \to gHg^{-1}$ by $f(h) = ghg^{-1}$ for all $h \in H$. Let $h, h' \in H$. If $h = h'$, then $ghg^{-1} = gh'g^{-1}$, i.e., f is well defined. Also, $ghg^{-1} \in gHg^{-1}$. Thus, f is a function of H into gHg^{-1}. Suppose $f(h) = f(h')$. Then $ghg^{-1} = gh'g^{-1}$. From this it follows that $h = h'$. This shows that f is one-one. To show f is onto gHg^{-1}, let $a \in gHg^{-1}$. Then $a = gbg^{-1} = f(b)$ for some $b \in H$, namely, $b = g^{-1}ag$. Thus, f is onto gHg^{-1}.

♢ **Exercise 2** Prove that S_n is generated by $\{(1\ 2), (1\ 3), (1\ 4), \ldots, (1\ n)\}$.

Solution Let π be any permutation in S_n. Then π is a product of transpositions. Thus, it is sufficient to show that if $(i\ j)$ is any transposition in S_n, $i < j$, then

$$(i\ j) \in \langle (1\ 2), (1\ 3), (1\ 4), \ldots, (1\ n) \rangle.$$

This follows from the fact that $(i\ j) = (1\ i) \circ (1\ j) \circ (1\ i)$. Hence, S_n is generated by $\{(1\ 2), (1\ 3), (1\ 4), \ldots, (1\ n)\}$.

[1] We show the existence of such groups in Chapter 7.

♢ **Exercise 3** Find all subgroups of $(\mathbb{Z}, +)$.

Solution Let H be a subgroup of $\mathbb{Z}$. Suppose $H \neq \{0\}$. Let a be a nonzero element of H. Then $-a \in H$. Since either a or $-a$ is a positive integer, H contains a positive integer. With the help of the principle of well-ordering, we can show that H contains a smallest positive integer. Let a be the smallest positive integer in H. We claim that $H = \{na \mid n \in \mathbb{Z}\}$.

Now $na \in H$ for all $n \in \mathbb{Z}$ and so $\{na \mid n \in \mathbb{Z}\} \subseteq H$. On the other hand, let $b \in H$. By the division algorithm, there exist c and r in $\mathbb{Z}$ such that $b = ca + r$, where $0 \leq r < a$. Suppose $r \neq 0$. Then $r = b - ca \in H$. Thus, H contains a positive integer smaller than a, a contradiction. Hence, $r = 0$ and so $b = ca \in \{na \mid n \in \mathbb{Z}\}$. This implies that $H \subseteq \{na \mid n \in \mathbb{Z}\}$. Thus, $H = \{na \mid n \in \mathbb{Z}\}$ for some $a \in \mathbb{Z}$. Also, for all $n \in \mathbb{Z}$, the set $T = \{nm \mid m \in \mathbb{Z}\} = n\mathbb{Z}$ is a subgroup of $\mathbb{Z}$. Hence, $n\mathbb{Z}$, $n = 0, 1, 2, \ldots$ are the subgroups of $\mathbb{Z}$.

Exercises

1. Prove that H is a subgroup of the group G, where
   (a) $H = \{[0], [2], [4], [6], [8], [10]\}$, $G = \mathbb{Z}_{12}$,
   (b) $H = \{[0], [3], [6], [9]\}$, $G = \mathbb{Z}_{12}$
   and where the group operation under consideration is $+_{12}$.

2. Let $GL(2, \mathbb{R})$ denote the group of all nonsingular 2×2 matrices over $\mathbb{R}$. Show that each of the following sets is a subgroup of $GL(2, \mathbb{R})$.
   (a) $S = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a, b, c, d \in \mathbb{R} \text{ and } ad - bc = 1 \right\}$.
   (b) $S = \left\{ \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} \mid a, b, c, d \in \mathbb{R} \text{ and } a \neq 0 \right\}$.
   (c) $S = \left\{ \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \mid a, b, c, d \in \mathbb{R} \text{ and either } a \text{ or } b \text{ is nonzero} \right\}$.
   (d) $S = \left\{ \begin{bmatrix} a & b \\ 0 & d \end{bmatrix} \mid a, b, c, d \in \mathbb{R} \text{ and } ad \neq 0 \right\}$.
   (e) $S = \left\{ \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \mid a, b \in \mathbb{R} \text{ and } a^2 + b^2 \neq 0 \right\}$.
   (f) $S = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a, b, c, d \in \mathbb{Z} \text{ and } ad - bc = 1 \right\}$

3. Show that the set $H = \{a + bi \in \mathbb{C}^* \mid a^2 + b^2 = 1\}$ is a subgroup of $(\mathbb{C}^*, \cdot)$, where $\cdot$ is the multiplication operation of complex numbers.

4. Let $G = \{(a, b) \mid a, b \in \mathbb{R}, b \neq 0\}$. Prove that $(G, *)$ is a noncommutative group under the binary operation $(a, b) * (c, d) = (a + bc, bd)$ for all $(a, b), (c, d) \in G$.
   (a) Let $H = \{(a, b) \in G \mid a = 0\}$. Show that H is a subgroup of G.
   (b) Let $K = \{(a, b) \in G \mid b > 0\}$. Show that K is a subgroup of G.
   (c) Let $T = \{(a, b) \in G \mid b = 1\}$. Show that T is a subgroup of G.
   (d) Find all elements of order 2 in G.

5. In S_3, determine the set $T = \{x \in S_3 \mid x^2 = e\}$. Is T a subgroup of S_3?

6. Determine the subgroup $\langle 4, 6 \rangle$ in $(\mathbb{Z}, +)$.

7. In $(\mathbb{Z}, +)$, determine the subgroup generated by $\{4, 5\}$.

8. List the elements of the following subgroups.
   (a) $\left\langle \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} \right\rangle$ in S_4.
   (b) $\langle h, v \rangle$ in the symmetries of the square.

9. Let $a = (1\ 2\ 3\ 4)$ and $b = (2\ 4) \in S_4$.

(a) Find $\circ(a)$ and $\circ(b)$.

(b) Show that $ba = a^3b = a^{-1}b$.

(c) Find $H = \langle a, b\rangle$ in S_4.

(d) Find $|H|$.

10. Let G be a group generated by a, b such that $\circ(b) = 2$, $\circ(a) = 6$, and $(ab)^2 = e$. Show that

(a) $aba = b$,

(b) $(a^2b)^2 = e$,

(c) $ba^2b = a^4$,

(d) $ba^3b = a^3$.

11. Let G be a group. Prove that a nonempty subset H of G is a subgroup if and only if for all $a, b \in H$, $ab \in H$ and $a^{-1} \in H$.

12. Let G be a commutative group. Show that the set H of all elements of finite order is a subgroup of G.

13. Let G be a group and $a \in G$. Show that if a is the only element of order n in G, then $a \in Z(G)$.

14. Show that $Z(S_n) = \{e\}$ for all $n \geq 3$.

15. Let G be a group and $a \in G$. Let $C(a) = \{b \in G \mid ba = ab\}$. Prove that $C(a)$ is a subgroup of G and that $Z(G) = \cap_{a\in G} C(a)$. $C(a)$ is called the **centralizer** of a in G.

16. Prove that a group G cannot be written as the union of two proper subgroups.

17. Let G be a group and H be a nonempty subset of G.

(a) Show that if H is a subgroup of G, then $HH = H$.

(b) If H is finite and $HH \subseteq H$, prove that H is a subgroup of G.

(c) Give an example of a group G and a nonempty subset H of G such that $HH \subseteq H$, but H is not a subgroup of G.

18. Let H be a subgroup of a group G. Prove that $\langle H\rangle = H$.

19. If A and B are subgroups of a group G, prove that $A \cup B$ is a subgroup of G if and only if $A \subseteq B$ or $B \subseteq A$. If C is also a subgroup of G, does a similar necessary and sufficient condition hold for $A \cup B \cup C$ to be a subgroup of G?

20. Let G be a commutative group. If a and b are two distinct elements of G such that $\circ(a) = 2 = \circ(b)$, show that $|\langle a, b\rangle| = 4$.

21

(a) Prove that S_n is generated by $\{(1\ 2), (1\ 2\ 3\ \cdots\ n)\}$.

(b) Prove that S_n is generated by $\{(1\ 2), (2\ 3), (3\ 4), \ldots, (n-1\ n)\}$.

22 Show that $(\mathbb{Q}, +)$ is not finitely generated.

23 Let G be a group. Prove that if G is finite, then G has finitely many subgroups.

24 Does there exist an infinite group with only a finite number of subgroups?

25 For the following statements, write the proof if the statement is true; otherwise, give a counterexample.

(a) All nontrivial subgroups of $(\mathbb{Z}, +)$ are infinite groups.

(b) If A, B, and C are subgroups of a group G such that $A \cup B \subseteq C$, then $ABC \subseteq C$.

(c) If G is a noncommutative group, then $Z(G) = \{e\}$.

(d) Let G be a group. If H is a nonempty subset of G such that $a^{-1} \in H$ for all $a \in H$, then H is a subgroup of G.

(e) There exists a proper subgroup A of $(\mathbb{Z}, +)$ such that A contains both $2\mathbb{Z}$ and $3\mathbb{Z}$.

(f) If H is a subgroup of $(\mathbb{Q}, +)$ such that $\mathbb{Z} \subset H$, then $H = \mathbb{Q}$.

(g) If H is a subgroup of $(\mathbb{Q}^*, \cdot)$ such that $\mathbb{Z}\backslash\{0\} \subseteq H$, then $H = \mathbb{Q}^*$.

4.2 Cyclic Groups

In the previous section, we introduced the notion of a subgroup generated by a set. Groups that are generated by a single element, called cyclic groups, are of special importance. As we shall see throughout the text, these groups play an important role in studying the structure of a group. In fact, all of Chapter 9 revolves around these groups. Cyclic groups are easier to study than any other group. They have special properties, some of which we will discover in this section.

Definition 4.2.1 *A group G is called a* ***cyclic group*** *if there exists $a \in G$ such that*

$$G = \langle a \rangle .$$

We recall that $\langle a \rangle$ in Definition 4.2.1 is the set $\{a^n \mid n \in \mathbb{Z}\}$ (Corollary 4.1.14)**.**

Let $G = \langle a \rangle$ be a cyclic group and $b, c \in G$. Then $b = a^n$ and $c = a^m$ for some $n, m \in \mathbb{Z}$. Now

$$bc = a^n a^m = a^{n+m} = a^{m+n} = a^m a^n = cb.$$

This shows that G is commutative. Hence, every cyclic group is commutative. We record this result in the following theorem.

Theorem 4.2.2 *Every cyclic group is commutative.*

Example 4.2.3 *(i)* $(\mathbb{Z}, +)$ *is a cyclic group because* $\mathbb{Z} = \langle 1 \rangle$.
(ii) $(\{na \mid n \in \mathbb{Z}\}, +)$ *(Example* 2.1.8*) is a cyclic group, where a is any fixed element of* $\mathbb{Z}$.
(iii) $(\mathbb{Z}_n, +_n)$ *is a cyclic group because* $\mathbb{Z}_n = \langle [1] \rangle$.

Example 4.2.4 *Let a be a symbol and n a positive integer. Define $*$ by means of the following operation table.*

$*$	a^0	a^1	a^2	$\cdots$	a^{n-2}	a^{n-1}
a^0	a^0	a^1	a^2	$\cdots$	a^{n-2}	a^{n-1}
a^1	a^1	a^2	a^3	$\cdots$	a^{n-1}	a^0
a^2	a^2	a^3	a^4	$\cdots$	a^0	a^1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a^{n-2}	a^{n-2}	a^{n-1}	a^0	$\cdots$	a^{n-4}	a^{n-3}
a^{n-1}	a^{n-1}	a^0	a^1	$\cdots$	a^{n-3}	a^{n-2}

Then $(\{a^0, a^1, \ldots, a^{n-1}\}, *)$ *is a cyclic group generated by* a^1.

Example 4.2.5 *Consider the set $G = \{e, a, b, c\}$. Define $*$ on G by means of the following operation table.*

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

From the multiplication table, it follows that $(G, *)$ is a commutative group. However, G is not a cyclic group because

$$\langle e \rangle = \{e\}, \langle a \rangle = \{e, a\}, \langle b \rangle = \{e, b\}, \text{ and } \langle c \rangle = \{e, c\}$$

and each of these subgroups is properly contained in G. G is known as the ***Klein 4-group****.*

The next theorem gives the exact description of a finite cyclic group.

Theorem 4.2.6 *Let $\langle a \rangle$ be a finite cyclic group of order n. Then* $\langle a \rangle = \{e, a, a^2, \ldots, a^{n-1}\}$.

Proof. By Corollary 4.1.14, $\langle a\rangle = \{a^i \mid i \in \mathbb{Z}\}$. Because $\langle a\rangle$ is finite, there exist $i, j \in \mathbb{Z}$ $(j > i)$ such that $a^i = a^j$. Thus, $a^{j-i} = e$ and $j - i$ is positive. Let m be the smallest positive integer such that $a^m = e$. Then for all integers i, j such that $0 \le i < j < m$, $a^i \neq a^j$ otherwise $a^{j-i} = e$ for some $0 \le i < j < m$, which contradicts the minimality of m. Hence, the elements of the set $S = \{e, a, a^2, \ldots, a^{m-1}\}$ are distinct. Clearly $S \subseteq \langle a\rangle$. Let $a^k \in \langle a\rangle$. By the division algorithm, there exist integers q, r such that $k = qm + r$, $0 \le r < m$. Thus, $a^k = a^{qm+r} = (a^m)^q a^r = ea^r = a^r \in S$. Therefore, $\langle a\rangle \subseteq S$. Thus, $S = \langle a\rangle$. Since the elements of S are distinct and $\langle a\rangle$ has order n, it must be the case that $m = n$. ■

The following corollaries are immediate from the proof of Theorem 4.2.6. We omit the proofs.

Corollary 4.2.7 *Let $\langle a\rangle$ be a finite cyclic group. Then $\circ(a) = |\langle a\rangle|$.* ■

Corollary 4.2.8 *A finite group G is a cyclic group if and only if there exists an element $a \in G$ such that $\circ(a) = |G|$.* ■

As stated in the beginning of this section, cyclic groups have special properties. We now proceed to discover some of these properties. Subgroups of a cyclic group are themselves cyclic; this is proved in the next theorem.

Theorem 4.2.9 *Every subgroup of a cyclic group is cyclic.*

Proof. Let H be a subgroup of a cyclic group $G = \langle a\rangle$. If $H = \{e\}$, then $H = \langle e\rangle$, so H is cyclic. Suppose $\{e\} \subset H$. Then there exists $b \in H$ such that $b \neq e$. Since $b \in G$, we have $b = a^m$ for some integer m. Thus, $m \neq 0$ since $b \neq e$. Since H is a group, $a^{-m} = b^{-1} \in H$. Now either m or $-m$ is positive. Therefore, H contains at least one element which is a positive power of a. Let n be the smallest positive integer such that $a^n \in H$. We now show that $H = \langle a^n\rangle$.

Now $a^n \in H$, so we must have $\langle a^n\rangle \subseteq H$. Let $h \in H$. Then $h = a^k$ for some integer k. By the division algorithm, there exist integers q, r such that $k = nq+r, 0 \le r < n$. Since a^n and $a^k \in H$, we have $a^r = a^{k-nq} = a^k(a^n)^{-q} \in H$. However, if $r > 0$, we contradict the minimality of n. Therefore, $r = 0$ so that $a^k = (a^n)^q \in \langle a^n\rangle$. Hence, $H \subseteq \langle a^n\rangle$, so $H = \langle a^n\rangle$. Thus, H is cyclic. ■

Corollary 4.2.10 *Let $G = \langle a\rangle$ be a cyclic group of order m, $m > 1$, and H be a proper subgroup of G. Then $H = \langle a^k\rangle$ for some integer k such that k divides m and $k > 1$. Furthermore, $|H|$ divides m.*

Proof. If $H = \{e\}$, then $H = \langle a^m\rangle$. Suppose that $H \neq \{e\}$. Let k be the smallest positive integer such that $a^k \in H$. Then $H = \langle a^k\rangle$. Now there exist integers q and r such that $m = qk + r$, where $0 \le r < k$, and

$$a^r = a^{m-qk} = a^m a^{-qk} = a^{-qk} = ((a^k)^{-1})^q \in H.$$

The minimality of k implies that $r = 0$. Hence, $m = qk$ and so k divides m. Since $H \neq G$, $k > 1$. Next, we show that $|H|$ divides m. By Theorem 2.1.46(ii), $\circ(a^k) = \frac{m}{\gcd(m,k)} = \frac{m}{k} = q$. As a result Corollary 4.2.7 implies that

$$|H| = \circ(a^k) = q.$$

Since $m = qk$, we have $q \mid m$, i.e., $|H|$ divides m. ■

By Corollary 4.2.10, if G is a finite cyclic group and H is a subgroup of G, then $|H|$ divides $|G|$. This is a special case of a more general result, called Lagrange's theorem, which we will prove in the next section.

Let $G = \langle a\rangle$ be an infinite cyclic group. Then $\circ(a)$ is infinite and this implies that $\circ(a^k)$ is infinite for any nonzero integer k. Thus, the order of any nonidentity element of G is infinite. Let H be a nontrivial subgroup of G. Then H is cyclic. Let $H = \langle b\rangle$. Then $b \neq e$ and $b \in G$ and so $\circ(b)$ is infinite. This in turn shows that $|H|$ is infinite. Thus, every nontrivial subgroup of an infinite cyclic group is infinite.

Now let $G = \langle a\rangle$ be a finite cyclic group of order n and H be a proper subgroup of G. Then by Corollary 4.2.10, $|H|$ divides $|G|$. If $H = \{e\}$, then $|H| = 1$ and if $H = G$, then $|H| = |G|$ and so $|H|$ divides $|G|$. Thus, the order of every subgroup of G divides the order of G. The following theorem shows that the converse of this result is also true for finite cyclic groups.

Theorem 4.2.11 *Let G be a finite cyclic group of order m. Then for every positive divisor d of m, there exists a unique subgroup of G of order d.*

Proof. Let $G = \langle a \rangle$ and d be a positive divisor of m. Becuase $d \mid m$, there exists $k \in \mathbb{Z}$ such that $m = kd$. Now $a^k \in G$ and by Theorem 2.1.46(ii),

$$\circ(a^k) = \frac{\circ(a)}{\gcd(k, m)} = \frac{m}{k} = d.$$

Let $H = \langle a^k \rangle$. Then $|H| = \circ(a^k) = d$. Thus, G has a subgroup of order d. Next, we establish that H is unique.

Let K be a subgroup of order d. Let t be the smallest positive integer such that $a^t \in K$. Then $K = \langle a^t \rangle$. Because K is of order d, $\circ(a^t) = d$ by Corollary 4.2.7. But $\circ(a^t) = \frac{m}{\gcd(t,m)}$ by Theorem 2.1.46(ii). Hence, $d = \frac{m}{\gcd(t,m)}$, which implies that $\gcd(t, m) = \frac{m}{d} = k$. This shows that $k \mid t$. Let $t = kl$ for some $l \in \mathbb{Z}$. Now $a^t = a^{kl} = (a^k)^l \in H$. Hence, $K \subseteq H$. Since $|K| = |H|$ and H and K are finite, we have $H = K$. Thus, there exists a unique subgroup of order d. ■

Worked-Out Exercises

♢ **Exercise 1** $(\mathbb{Q}, +)$is not cyclic.

Solution Suppose $\mathbb{Q}$ is cyclic. Then $\mathbb{Q}=\left\langle \frac{p}{q} \right\rangle$ for some $\frac{p}{q} \in \mathbb{Q}$, where p and q are relatively prime. Since $\frac{p}{2q} \in \mathbb{Q}$, there exists $n \in \mathbb{Z}$, $n \neq 0$ such that $\frac{p}{2q} = n\frac{p}{q}$ by Corollary 4.1.14. This implies that $\frac{1}{2} = n \in \mathbb{Z}$, which is a contradiction. Thus, $\mathbb{Q}$ is not cyclic.

Exercise 2 Let G be a group such that $|G| = mn$, $m > 1$, $n > 1$. Show that G has a nontrivial subgroup.

Solution First suppose that G is cyclic. Let $G = \langle a \rangle$. Then $\circ(a) = mn$. Clearly $\circ(a^m) = n$. Let $H = \langle a^m \rangle$. Then H is a nontrivial subgroup of G. Now suppose that G is not cyclic. Then for all $a \in G$, $\circ(a) < mn$ by Exercise 26 (page 56). Let $e \neq a \in G$ and let $H = \langle a \rangle$. Then H is a nontrivial subgroup of G.

♢ **Exercise 3** Let G be an infinite cyclic group generated by a. Show that

(a) $a^r = a^t$ if and only if $r = t$, where $r, t \in \mathbb{Z}$,

(b) G has exactly two generators.

Solution (a) Suppose $a^r = a^t$ and $r \neq t$. Let $r > t$. Then $a^{r-t} = e$. Thus, $\circ(a)$ is finite, say, $\circ(a) = n$. Then $G = \{e, a, \ldots, a^{n-1}\}$, which is a contradiction since G is an infinite group. The converse is straightforward.

(b) Let $G = \langle b \rangle$ for some $b \in G$. Since $a \in G = \langle b \rangle$ and $b \in G = \langle a \rangle$, $a = b^r$ and $b = a^t$ for some $r, t \in \mathbb{Z}$. Thus, $a = b^r = (a^t)^r = a^{rt}$. Hence, by (a), $rt = 1$. This implies that either $r = 1 = t$ or $r = -1 = t$. Thus, either $b = a$ or $b = a^{-1}$. Now from (a), $a \neq a^{-1}$. Therefore, G has exactly two generators.

Exercise 4 (a) Let $G = \langle a \rangle$ be a finite cyclic group of order n. Show that a^k is a generator of G if and only if $\gcd(k, n) = 1$, where k is a positive integer.

(b) Find all generators of $\mathbb{Z}_{10}$.

Solution (a) Suppose a^k is a generator of G. Since $|G| = n$, $\circ(a^k) = n$. But $\circ(a^k) = \frac{n}{\gcd(k,n)}$. Hence, $\frac{n}{\gcd(k,n)} = n$. Thus, $\gcd(k, n) = 1$. Conversely, suppose that $\gcd(k, n) = 1$. Then $\circ(a^k) = \frac{n}{\gcd(k,n)} = n$. Hence, $\left|\langle a^k \rangle\right| = n$. Since $\langle a^k \rangle \subseteq G$ and $|G| = n$, $G = \langle a^k \rangle$.

(b) Now $\mathbb{Z}_{10} = \langle [1] \rangle$ and $|\mathbb{Z}_{10}| = 10$. By (a), $k[1]$ is a generator if and only if $\gcd(k, 10) = 1$, where $1 \leq k \leq 10$. Now if $k = 1, 3, 7$, or 9, then $\gcd(k, 10) = 1$. Thus, the generators of $\mathbb{Z}_{10}$ are $1[1] = [1]$, $3[1] = [3]$, $7[1] = [7]$ and $9[1] = [9]$.

Exercises

1. Let $G = \langle a \rangle$ be a cyclic group of order 30. Determine the following subgroups.
   (a) $\langle a^5 \rangle$.
   (b) $\langle a^2 \rangle$.
2. Let G be a cyclic group of order 30. Find the number of elements of order 6 in G and also find the number of elements of order 5 in G.
3. Prove that 1 and -1 are the only generators of $\mathbb{Z}$.
4. (a) Show that $(\mathbb{R}, +)$ is not cyclic.

(b) Show that $(\mathbb{Q}^*, \cdot)$ is not cyclic.

(c) Show that $(\mathbb{R}^*, \cdot)$ is not cyclic.

5. If G is a cyclic group of order n, show that the number of generators of G is $\phi(n)$, where ϕ is the Euler ϕ-function.

6. Show that every proper subgroup of S_3 is cyclic.

7. Give an example of a noncyclic Abelian group all of whose proper subgroups are cyclic.

8. Let G be a group. Suppose that G has at most two nontrivial subgroups. Show that G is cyclic.

9. Let G be a finite group. Show that if G has exactly one nontrivial subgroup, then order of G is p^2 for some prime p.

10. Let G be a noncommutative group. Show that G has a nontrivial subgroup.

11. Give an example of an infinite group of order ≥ 2 which contains a nontrivial finite cyclic group.

12. Show that there are cyclic subgroups of order $1, 2, 3$, and 4 in S_4, but S_4 does not contain any cyclic subgroup of order ≥ 5.

13. For the following statements, write the proof if the statement is true; otherwise, give a counterexample.

(a) For every positive integer n, there exists a cyclic group of order n.

(b) Every proper subgroup of A_4 is cyclic.

(c) A_3 is a cyclic group.

(d) A_4 is a cyclic group.

(e) All proper subgroups of $(\mathbb{R}, +)$ are cyclic.

4.3 Lagrange's Theorem

In the last section, we noted that the order of a subgroup of a finite cyclic group divides the order of the group (Corollary 4.2.10). We also remarked that this is a special case of a general result, called Lagrange's theorem, i.e., the order of a subgroup of a finite group divides the order of the group. Lagrange proved this result in 1770, long before the creation of group theory, while working on the permutations of the roots of a polynomial equation. Lagrange's theorem is a basic theorem of finite group theory and is considered by some to be the most important result in finite group theory. In this section, we prove this result. We begin with the following definition.

Definition 4.3.1 *Let H be a subgroup of a group G and $a \in G$. The sets $aH = \{ah \mid h \in H\}$ and $Ha = \{ha \mid h \in H\}$ are called the **left** and **right cosets** of H in G, respectively. The element a is called a representative of aH and Ha.*

If G is commutative, then of course $aH = Ha$. Observe that $eH = H = He$ and that $a = ae \in aH$ and $a = ea \in Ha$.

Example 4.3.2 *Consider the symmetric group S_3 (Example* 3.1.8*). Then*

$$H = \left\{e, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}\right\}$$

and

$$H' = \left\{e, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}\right\}$$

are subgroups of S_3. We now compute the left and right cosets of H in S_3. The left cosets of H in S_3 are

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

and

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} H = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} H = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} H = \left\{\begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}\right\}$$

and the right cosets of H in S_3 are

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

and

$$H\begin{pmatrix}1&2&3\\1&3&2\end{pmatrix} = H\begin{pmatrix}1&2&3\\3&2&1\end{pmatrix} = H\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix} = \left\{\begin{pmatrix}1&2&3\\1&3&2\end{pmatrix}, \begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}, \begin{pmatrix}1&2&3\\3&2&1\end{pmatrix}\right\}.$$

Thus, for all $a \in S_3$, $aH = Ha$.

Next, we compute the left and right cosets of H' in S_3. The left cosets of H' in S_3 are

$$\begin{pmatrix}1&2&3\\1&2&3\end{pmatrix}H' = \begin{pmatrix}1&2&3\\1&3&2\end{pmatrix}H' = H',$$

$$\begin{pmatrix}1&2&3\\3&2&1\end{pmatrix}H' = \begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}H' = \left\{\begin{pmatrix}1&2&3\\3&2&1\end{pmatrix}, \begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}\right\},$$

and

$$\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}H' = \begin{pmatrix}1&2&3\\2&3&1\end{pmatrix}H' = \left\{\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}, \begin{pmatrix}1&2&3\\2&3&1\end{pmatrix}\right\}$$

and the right cosets of H' in S_3 are

$$H'\begin{pmatrix}1&2&3\\1&2&3\end{pmatrix} = H'\begin{pmatrix}1&2&3\\1&3&2\end{pmatrix} = H',$$

$$H'\begin{pmatrix}1&2&3\\3&2&1\end{pmatrix} = H'\begin{pmatrix}1&2&3\\2&3&1\end{pmatrix} = \left\{\begin{pmatrix}1&2&3\\3&2&1\end{pmatrix}, \begin{pmatrix}1&2&3\\2&3&1\end{pmatrix}\right\},$$

and

$$H'\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix} = H'\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix} = \left\{\begin{pmatrix}1&2&3\\2&1&3\end{pmatrix}, \begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}\right\}.$$

We see that

$$\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}H' \neq H'\begin{pmatrix}1&2&3\\3&1&2\end{pmatrix}.$$

Thus, the left and right cosets of H' in S_3 are not the same.

There are some interesting phenomena happening in the above example. We see that all left and right cosets of H in S_3 have the same number of elements, namely, 3; that there are the same number of distinct left cosets of H in S_3 as of right cosets, namely, 2; that the set of all left cosets and the set of all right cosets form partitions of S_3; and, finally, that $3 \cdot 2$ equals the order of S_3. Similar statements hold for the subgroup H'. We show, in the results to follow, that these phenomena hold in general.

In the next few theorems, we prove some properties of left and right cosets of a subgroup which will eventually lead us to the proof of Lagrange's theorem. The following theorem tells us when two left (right) cosets are equal. It is a result that is used often in the study of groups.

Theorem 4.3.3 *Let H be a subgroup of a group G and $a, b \in G$. Then*
(i) $aH = bH$ if and only if $b^{-1}a \in H$.
(ii) $Ha = Hb$ if and only if $ab^{-1} \in H$.

Proof. (i) Suppose $aH = bH$. Since $a \in aH$ and $aH = bH$, there exists $h' \in H$ such that $a = bh'$. This implies that $b^{-1}a = h' \in H$.

Conversely, suppose $b^{-1}a \in H$. Then there exists $h' \in H$ such that $b^{-1}a = h'$, i.e., $a = bh'$. Let $ah \in aH$. Then $ah = bh'h \in bH$. This implies that $aH \subseteq bH$. Next, we show that $bH \subseteq aH$. Now $b^{-1}a = h'$ implies that $ah'^{-1} = b$. Let $bh \in bH$. Then $bh = ah'^{-1}h \in aH$. Hence, $bH \subseteq aH$. Consequently, $aH = bH$.

(ii) The proof is similar to (i). We leave it as an exercise. ■

Theorem 4.3.4 *Let H be a subgroup of a group G. Then for all a, $b \in G$, either $aH = bH$ or $aH \cap bH = \emptyset$ (i.e., two left cosets are either equal or they are disjoint).*

Proof. Let $a, b \in G$. Suppose that $aH \cap bH \neq \emptyset$. We wish to show that $aH = bH$. Since $aH \cap bH \neq \emptyset$, there exists $c \in aH \cap bH$. Hence, $c \in aH$ and $c \in bH$ and so there exist h_1, $h_2 \in H$ such that $c = ah_1$ and $c = bh_2$. Thus, $ah_1 = bh_2$ and from this, it follows that $b^{-1}a = h_2h_1^{-1}$. Therefore, $b^{-1}a \in H$. By Theorem 4.3.3(i), $aH = bH$. ■

Corollary 4.3.5 *Let H be a subgroup of a group G. Then $\{aH \mid a \in G\}$ forms a partition of G.*

Proof. Let $\mathcal{P} = \{aH \mid a \in G\}$, i.e., $\mathcal{P}$ is the set of all left cosets of H in G. By Theorem 4.3.4, for all $aH, bH \in \mathcal{P}$, either $aH = bH$ or $aH \cap bH = \emptyset$. Thus, $\mathcal{P}$ satisfies (i) of Definition 1.3.14. Since $aH \subseteq G$ for all $a \in G$, $\cup_{aH \in \mathcal{P}} aH \subseteq G$. If $a \in G$, then $a \in aH \subseteq \cup_{aH \in \mathcal{P}} aH$. Therefore, $G \subseteq \cup_{aH \in \mathcal{P}} aH$. Hence, $G = \cup_{aH \in \mathcal{P}} aH$. This shows that $\mathcal{P}$ satisfies (ii) of Definition 1.3.14. Consequently, $\mathcal{P}$ is a partition of G. ■

Theorem 4.3.6 *Let H be a subgroup of a group G. Then the elements of H are in one-one correspondence with the elements of any left (right) coset of H in G.*

Proof. Let a be any element of G and aH be a left coset of H in G. To show that the elements of H are in one-one correspondence with the elements of aH, we show that there exists a one-one function of H onto aH. Define $f : H \to aH$ by $f(h) = ah$ for all $h \in H$. Let $h, h_1 \in H$. If $h = h_1$, then $ah = ah_1$, i.e., $f(h) = f(h_1)$. Hence, f is well defined. Suppose $f(h) = f(h_1)$. Then $ah = ah_1$ and this implies that $h = h_1$. Thus, f is a one-one function. To show f is onto aH, let $ah \in aH$, where $h \in H$. Then $ah = f(h)$. Hence, f maps H onto aH. Similarly, we can show that the elements of H are in one-one correspondence with the elements of Ha. ■

The following corollary is immediate from Theorem 4.3.6.

Corollary 4.3.7 *Let H be a subgroup of a group G. Then for all $a \in G$, $|H| = |aH| = |Ha|$.* ■

The next theorem says that there are the same number of left cosets as right cosets.

Theorem 4.3.8 *Let H be a subgroup of a group G. Then there is a one-one correspondence of the set of all left cosets of H in G onto the set of all right cosets of H in G.*

Proof. Let $\mathcal{L} = \{aH \mid a \in G\}$ be the set of all left cosets of H in G and $\mathcal{R} = \{Ha \mid a \in G\}$ be the set of all right cosets of H in G. To establish a one-one correspondence between the elements of $\mathcal{L}$ and $\mathcal{R}$, we need to show the existence of a one-one function of $\mathcal{L}$ onto $\mathcal{R}$.

Define $f : \mathcal{L} \to \mathcal{R}$ by

$$f(aH) = Ha^{-1}$$

for all $aH \in \mathcal{L}$. First note that $Ha^{-1} \in \mathcal{R}$ for all $a \in G$. Let $aH, bH \in \mathcal{L}$. Suppose $aH = bH$. Then by Theorem 4.3.3(i), $b^{-1}a \in H$. This implies that $b^{-1}(a^{-1})^{-1} = b^{-1}a \in H$ and so by Theorem 4.3.3(ii), $Hb^{-1} = Ha^{-1}$. Thus, $f(bH) = f(aH)$. Hence, f is well defined. To show f is one-one, suppose $f(aH) = f(bH)$. Then $Ha^{-1} = Hb^{-1}$ and so $a^{-1}(b^{-1})^{-1} \in H$ by Theorem 4.3.3(ii), i.e., $a^{-1}b \in H$. Therefore, $b^{-1}a = (a^{-1}b)^{-1} \in H$ and so $aH = bH$. Hence, f is one-one. Since for all $Ha \in \mathcal{R}$, $Ha = H(a^{-1})^{-1} = f(a^{-1}H)$ and $a^{-1}H \in \mathcal{L}$, it follows that f is onto $\mathcal{R}$. Thus, f is a one-one function from $\mathcal{L}$ onto $\mathcal{R}$. ■

Definition 4.3.9 *Let H be a subgroup of a group G. Then the number of distinct left (or right) cosets, written $[G : H]$, of H in G is called the **index** of H in G.*

By Theorem 4.3.8, the number of left cosets and the number of right cosets of a subgroup H of a group G are the same. Thus, $[G : H]$ is well defined.

If G is finite, then of course $[G : H]$ is finite. The following example is one, where G is infinite and $[G : H]$ is finite.

Example 4.3.10 *Let n be a fixed positive integer. Consider the cyclic subgroup $(\langle n \rangle, +)$ of $(\mathbb{Z}, +)$. Let $k + \langle n \rangle$ be a left coset of $\langle n \rangle$ in $\mathbb{Z}$. By the division algorithm, there exist integers q and r such that $k = qn + r$, where $0 \leq r < n$. Then $k - r = qn \in \langle n \rangle$ and so $k + \langle n \rangle = r + \langle n \rangle$ by Theorem 4.3.3. Suppose $i + \langle n \rangle = j + \langle n \rangle$, where $0 \leq i, j < n$. Then $i - j \in \langle n \rangle$ by Theorem 4.3.3. This implies that $n \mid (i - j)$ and so we must have $i - j = 0$ or $i = j$ since $0 \leq i, j < n$. Thus, the distinct left cosets of $\langle n \rangle$ in $\mathbb{Z}$ are $0 + \langle n \rangle, 1 + \langle n \rangle, \ldots, n - 1 + \langle n \rangle$.*

We are now ready to prove Lagrange's theorem. It is interesting to note that Lagrange proved the result for the symmetric group S_n. Some credit Galois for proving the result in general.

Theorem 4.3.11 (Lagrange) *Let H be a subgroup of a finite group G. Then the order of H divides the order of G. In particular,*

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

Proof. Since G is a finite group, the number of left cosets of H in G is finite. Let $\{a_1H, a_2H, \ldots, a_rH\}$ be the set of all distinct left cosets of H in G. Then by Corollary 4.3.5, $G = \cup_{i=1}^{r} a_iH$ and $a_iH \cap a_jH = \emptyset$ for all $i \neq j$, $1 \leq i, j \leq r$. Hence, $[G : H] = r$ and

$$|G| = |a_1H| + |a_2H| + \cdots + |a_rH|\,.$$

By Corollary 4.3.7, $|H| = |a_iH|$ for all i, $1 \leq i \leq r$. Therefore,

$$\begin{aligned} |G| &= \underbrace{|H| + |H| + \cdots + |H|}_{r \text{ times}} \\ &= r\,|H| \\ &= [G : H]\,|H|\,. \end{aligned}$$

Thus, the order of H divides the order of G. ■

Corollary 4.3.12 *Let G be a group of finite order n. Then the order of any element a of G divides n and $a^n = e$.*

Proof. Let $a \in G$ and $\circ(a) = k$. Let $H = \langle a \rangle$. Then by Corollary 4.2.7, $|H| = |\langle a \rangle| = \circ(a) = k$. Hence, by Theorem 4.3.11, k divides n. Thus, there exists $q \in \mathbb{Z}$ such that $n = kq$. Hence, $a^n = a^{kq} = (a^k)^q = e^q = e$. ■

Let G be a finite group of order n and $a \in G$. Then $\circ(a)$ divides n by Corollary 4.3.12. Thus, to find $\circ(a)$, we only need to check a^k, where k is a positive divisor of n. For example, consider $\mathbb{Z}_{20}$ and $[6] \in \mathbb{Z}_{20}$. Now $|\mathbb{Z}_{20}| = 20$ and 1, 2, 4, 5, 10, and 20 are the only positive divisors of 20. Now $1[6] = [6] \neq [0]$, $2[6] = [12] \neq [0]$, $4[6] = [24] = [4] \neq [0]$, $5[6] = [30] = [10] \neq [0]$, and $10[6] = [60] = [0]$. Thus, $\circ([6]) = 10$. Hence, the above corollary can be used to find the order of an element in a finite group.

Corollary 4.3.13 *Let G be a group of prime order. Then G is cyclic.*

Proof. Since $|G| \geq 2$, there exists $a \in G$ such that $a \neq e$. Let $H = \langle a \rangle$. Then $\{e\} \subset H$ and $|H|$ divides $|G|$. But $|G|$ is prime and so $|H| = |G|$. Since $H \subseteq G$ and $|H| = |G|$, it follows that $G = H$. Therefore, G is cyclic. ■

Let H and K be subgroups of a group G. If either H or K is infinite, then, of course, HK is infinite. Suppose H and K are both finite. We know that HK need not be a subgroup of G. Thus, $|HK|$ need not divide $|G|$. However, with the help of Lagrange's theorem, we can determine $|HK|$. This is a very useful result and we will use it very effectively in this text. In the next theorem, we determine $|HK|$ when H and K are both finite.

Theorem 4.3.14 *Let H and K be finite subgroups of a group G. Then*

$$|HK| = \frac{|H|\,|K|}{|H \cap K|}.$$

Proof. Let us write $A = H \cap K$. Since H and K are subgroups of G, A is a subgroup of G and since $A \subseteq H$, A is also a subgroup of H. By Lagrange's theorem, $|A|$ divides $|H|$. Let $n = \frac{|H|}{|A|}$. Then $[H : A] = n$ and so A has n distinct left cosets in H. Let $\{x_1A, x_2A, \ldots, x_nA\}$ be the set of all distinct left cosets of A in H. Then $H = \cup_{i=1}^{n} x_iA$. Since $A \subseteq K$, it follows that

$$HK = (\cup_{i=1}^{n} x_iA)K = \cup_{i=1}^{n} x_iK.$$

We now show that $x_iK \cap x_jK = \emptyset$ if $i \neq j$. Suppose $x_iK \cap x_jK \neq \emptyset$ for some $i \neq j$. Then $x_jK = x_iK$. Thus, $x_i^{-1}x_j \in K$. Since $x_i^{-1}x_j \in H$, $x_i^{-1}x_j \in A$ and so $x_jA = x_iA$. This contradicts the assumption that $x_1A, \ldots, x_nA$ are all distinct left cosets. Hence, $x_1K, \ldots, x_nK$ are distinct left cosets of K. Also, $|K| = |x_iK|$ by Corollary 4.3.7 for all $i = 1, 2, \ldots, n$. Thus,

$$\begin{aligned} |HK| &= |x_1K| + \cdots + |x_nK| \\ &= \underbrace{|K| + \cdots + |K|}_{n \text{ times}} \\ &= n\,|K| \\ &= \frac{|H||K|}{|A|} \\ &= \frac{|H||K|}{|H \cap K|}. \end{aligned}$$

■

The following corollary is an immediate consequence of the above theorem.

Corollary 4.3.15 *Let H and K be finite subgroups of a group G such that $H \cap K = \{e\}$. Then*

$$|HK| = |H|\,|K|\,. \blacksquare$$

Let H and K be subgroups of a group G. If either H or K is infinite, then, of course, HK is infinite. Suppose H and K are both finite. We know that HK need not be a subgroup of G. Thus, $|HK|$ need not divide $|G|$. However, with the help of Lagrange's theorem, we can determine $|HK|$. This is a very useful result and we will use it very effectively in this text. In the next theorem, we determine $|HK|$ when H and K are both finite.

Theorem 4.3.16 *Let H and K be finite subgroups of a group G. Then*

$$|HK| = \frac{|H|\,|K|}{|H \cap K|}.$$

Proof. Let us write $A = H \cap K$. Since H and K are subgroups of G, A is a subgroup of G and since $A \subseteq H$, A is also a subgroup of H. By Lagrange's theorem, $|A|$ divides $|H|$. Let $n = \frac{|H|}{|A|}$. Then $[H : A] = n$ and so A has n distinct left cosets in H. Let $\{x_1A,\, x_2A, \ldots, x_nA\}$ be the set of all distinct left cosets of A in H. Then $H = \cup_{i=1}^{n} x_iA$. Since $A \subseteq K$, it follows that

$$HK = (\cup_{i=1}^{n} x_iA)K = \cup_{i=1}^{n} x_iK.$$

We now show that $x_iK \cap x_jK = \emptyset$ if $i \neq j$. Suppose $x_iK \cap x_jK \neq \emptyset$ for some $i \neq j$. Then $x_jK = x_iK$. Thus, $x_i^{-1}x_j \in K$. Since $x_i^{-1}x_j \in H$, $x_i^{-1}x_j \in A$ and so $x_jA = x_iA$. This contradicts the assumption that $x_1A, \ldots, x_nA$ are all distinct left cosets. Hence, $x_1K, \ldots, x_nK$ are distinct left cosets of K. Also, $|K| = |x_iK|$ by Corollary 4.3.7 for all $i = 1, 2, \ldots, n$. Thus,

$$\begin{aligned} |HK| &= |x_1K| + \cdots + |x_nK| \\ &= \underbrace{|K| + \cdots + |K|}_{n \text{ times}} \\ &= n\,|K| \\ &= \frac{|H||K|}{|A|} \\ &= \frac{|H||K|}{|H \cap K|}. \end{aligned}$$

■

The following corollary is an immediate consequence of the above theorem.

Corollary 4.3.17 *Let H and K be finite subgroups of a group G such that $H \cap K = \{e\}$. Then*

$$|HK| = |H|\,|K|\,. \blacksquare$$

The following is known as Fermat's little theorem.

Theorem 4.3.18 (Fermat) *Let p be a prime integer and a be an integer such that p does not divide a. Then p divides $a^{p-1} - 1$, i.e.,*

$$a^{p-1} \equiv_p 1.$$

Proof. Let $U_p = \mathbb{Z}_p \backslash \{0\}$. Then by Exercise 10 (page 55), U_p is a group. Also, by Exercise 9 (page 55), $|U_p| = p - 1$. Let a be an integer such that p does not divide a. Then $[a]$ is a nonzero element of $\mathbb{Z}_p$ and so $[a] \in U_p$. Thus, by Corollary 4.3.12, $[a]^{p-1} = [1]$, i.e., $[a^{p-1}] = [1]$. Hence, $a^{p-1} \equiv_p 1$ by Exercise 11 (page 23). ■

We now discuss Euler's generalization of Fermat's little theorem.

If p is a prime, then $\phi(p) = p - 1$. Therefore, Fermat's little theorem can be written as follows:

If a is a positive integer and p is a prime such that $\gcd(a, p) = 1$, then

$$a^{\phi(p)} \equiv_p 1.$$

Euler generalized this result from the case of a prime to an arbitrary integer n. The following theorem is the Euler's generalization of Fermat's Little Theorem. We leave the proof as an exercise.

Theorem 4.3.19 (Euler) *Let a and n be integers such that $n > 0$ and $\gcd(a, n) = 1$. Then*

$$a^{\phi(n)} \equiv_n 1.$$

Worked Out Exercises

♢ **Exercise 1** Let H be a subgroup of a group G. Show that for all $a \in G$, $aH = H$ if and only if $a \in H$.

Solution Let $a \in G$. Suppose $aH = H$. Then $a = ae \in aH = H$. Conversely, suppose that $a \in H$. Now for any $h \in H$, $ah \in H$. Hence, $aH \subseteq H$. Let $h \in H$. Then $a^{-1}h \in H$. Thus, $h = a(a^{-1}h) \in aH$. Therefore, $H \subseteq aH$, proving that $aH = H$.

♢ **Exercise 2** Let G be a noncyclic group of order p^2, p a prime integer. Show that the order of each nonidentity element is p.

Solution Let $g \in G$ and $g \neq e$. Now $\circ(g)$ divides $|G| = p^2$. Hence, $\circ(g) = 1, p$ or p^2. Since $g \neq e$, $\circ(g) \neq 1$. If $\circ(g) = p^2$, then G contains an element g such that $\circ(g) = |G|$ and this implies that G is cyclic, which contradicts the hypothesis. Hence, $\circ(g) = p$.

Exercise 3 Let $G = \{a, b, c, d\}$ be a group. Complete the following Cayley table for this group.

	a	b	c	d
a				
b				
c			b	
d		b		

Solution From the table, $c^2 = b$ and $db = b$. Now $db = b$ implies that $d = e$, the identity element of G. Since $c^2 = b \neq d$, $\circ(c) \neq 2$. Hence, $\circ(c) = 4$. Thus, G is a cyclic group generated by c. Then $G = \{e\ , c, c^2, c^3\}$. Since $d = e$ and $c^2 = b$, it follows that $c^3 = a$. Hence, the Cayley table is

	a	b	c	d
a	b	c	d	a
b	c	d	a	b
c	d	a	b	c
d	a	b	c	d

Exercise 4 Let G be a finite nontrivial group. Suppose for all $x \in G$, there exists $y \in G$ such that $x = y^2$. Prove that the order of G is odd and conversely.

Solution Suppose G is of odd order. Then $|G| = 2n + 1$ for some positive integer n and for all $x \in G$, $x^{2n+1} = e$. Now $x^{2n+1} = e$ implies $x = x^{-2n} = (x^{-n})^2 = y^2$, where $y = x^{-n}$. Conversely, suppose $|G|$ is not odd. Let $|G| = 2n$ and $x \in G$. Then there exists $y \in G$ such that $x = y^2$. Hence, $x^n = y^{2n} = e$. Thus, for all $x \in G$, $x^n = e$. Suppose n is odd, say, $n = 2m + 1$. Then $x^{2m+1} = e$ for all $x \in G$. By Worked-Out Exercise 5 (page 52), there exists $z \in G$ such that $z \neq e$ and $z^2 = e$ since $|G|$ is even. Hence, $e = z^{2m+1} = zz^{2m} = z(z^2)^m = ze = z$, which is a contradiction. So n is even, say, $n = 2m$. Then $x^{2m} = e$ for all $x \in G$. As before, we can show that $x^m = e$ for all $x \in G$ and m is even. Continuing in this way, we can conclude that $x^2 = e$ for all $x \in G$. Let $x \in G$. Then there exists $y \in G$ such that $x = y^2$. Therefore, $x = e$. Thus, $|G| = 1$, which is a contradiction. Consequently, G is of odd order.

♢ **Exercise 5** Let G be a group such that $|G| > 1$. Prove that G has only the trivial subgroups if and only if $|G|$ is prime.

Solution Let $|G| = p$, p a prime. Let H be a subgroup of G. Then $|H|$ divides $|G|$. This implies that $|H| = 1$ or p. Thus, $H = \{e\}$ or $H = G$. Conversely, suppose that G has only the trivial subgroups. Let $a \in G$ be such that $a \neq e$. Now $\langle a\rangle = \{a^n \mid a \in \mathbb{Z}\}$ is a cyclic subgroup of G and $\langle a\rangle \neq \{e\}$. Therefore, $G = \langle a\rangle$. If G is infinite, then $a^r \neq a^s$ for all $r, s \in \mathbb{Z}$, $r \neq s$. Hence, $\{a^{2n} \mid n \in \mathbb{Z}\}$ is a nontrivial subgroup of G, which is a contradiction. Thus, $|G|$ is a finite cyclic group of order, say, $m > 1$. Suppose m is not prime. Then $m = rs$ for some $r, s \in \mathbb{Z}$, $1 < r, s < m$. Since $r \mid |G|$ and G is cyclic, G has a cyclic subgroup H of order r. This contradicts the assumption that G has only the trivial subgroups. Hence, $|G|$ is prime.

♢ **Exercise 6** Let G be a group of order p^n, p a prime. Show that G contains an element of order p.

Solution: Let $a \in G$, $a \neq e$. Then $H = \langle a\rangle$ is a cyclic subgroup of G. Now $|H|$ divides $|G| = p^n$. Thus, $|H| = p^m$ for some $m \in \mathbb{Z}$, $0 < m \leq n$. Now H is a cyclic group of order p^m. Hence, for every divisor d of p^m, there exists a subgroup of order d. So for p, there exists a subgroup T of H such that $|T| = p$. By Corollary 4.3.13, there exists $b \in T$ such that $T = \langle b\rangle$ and b is of order p. Hence, G contains an element of order p.

Exercise 7 Let G be a finite commutative group such that G contains two distinct elements of order 2. Show that $|G|$ is a multiple of 4. Also, show that this result need not be true if G is not commutative.

Solution Let a and b be two distinct elements of order 2. Let $H = \{e, a\}$ and $K = \{e, b\}$. Now H and K are subgroups of G. Since G is commutative, $HK = \{e, a, b, ab\}$ is a subgroup of G of order 4. Now $|HK| = 4$ divides $|G|$. Thus, $|G|$ is a multiple of 4.

The symmetric group S_3 is noncommutative, (1 2) and (1 3) are elements of S_3, and each is of order 2. But 4 does not divide $|S_3| = 6$.

Exercise 8 Find all subgroups of S_3.

Solution $S_3 = \{e, (1\ 2), (1\ 3), (2\ 3), (1\ 2\ 3), (1\ 3\ 2)\}$. $\circ(1\ 2) = 2$, $\circ(1\ 3) = 2$, $\circ(2\ 3) = 2$, $\circ(1\ 2\ 3) = 3$, and $\circ(1\ 3\ 2) = 3$. Now $\{e\}$, $\{e, (1\ 2)\}$, $\{e, (1\ 3)\}$, $\{e, (2\ 3)\}$, $\{e, (1\ 2\ 3), (1\ 3\ 2)\}$, and S_3 are subgroups of S_3. Let H be a subgroup of S_3. Now $|H|$ divides $|G|$. Thus, $|H| = 1, 2, 3$, or 6. If $|H| = 1$, then $H = \{e\}$. If $|H| = 6$, then $H = S_3$. If $|H| = 2$, then H is a cyclic group of order 2. Hence, H is one of $\{e, (1\ 2)\}$, $\{e, (1\ 3)\}$, $\{e, (2\ 3)\}$. Suppose $|H| = 3$. Then by Lagrange's theorem, H has no subgroup of order 2. Thus, (1 2), (1 3), (2 3) $\notin H$. Therefore, e, (1 2 3), (1 3 2) $\in H$. Also, $\{e, (1\ 2\ 3), (1\ 3\ 2)\}$ is a subgroup and so $H = \{e, (1\ 2\ 3), (1\ 3\ 2)\}$. Hence, $H_0 = \{e\}$, $H_1 = \{e, (1\ 2)\}$, $H_2 = \{e, (1\ 3)\}$, $H_3 = \{e, (2\ 3)\}$, $H_4 = \{e, (1\ 2\ 3), (1\ 3\ 2)\}$, and S_3 are the only subgroups of S_3.

Exercises

1. In S_3,
   (a) find all the right cosets of $H = \{e, (2\ 3)\}$,
   (b) find a subgroup B of G such that $H(1\ 2\ 3)$ is a left coset of B.
2. Find all right cosets of the subgroup $6\mathbb{Z}$ in the group $(\mathbb{Z}, +)$.
3. Let
$$H = \left\{e, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix}\right\},$$
   where e is the identity permutation. Show that H is a subgroup of S_4. List all the left and right cosets of H in S_4.
4. Let H denote the subgroup $\{r_{360}, h\}$ of the group of symmetries of the square. List all the left and right cosets of H in G.
5. Find all subgroups of the Klein 4-group.
6. Find all subgroups of order 4 in S_4.
7. Let $G = \{a, b, c, d\}$ be a group. Complete the following Cayley table for this group.

	d	a	b	c
d	d			
a		c	d	
b				
c				

8. Let G be a group and H and K be subgroups of G. Show that $(H \cap K)x = Hx \cap Kx$ for all $x \in G$.
9. Let G be a group and H and K be subgroups of G. Let $a, b \in G$. Show that either $Ha \cap Kb = \emptyset$ or $Ha \cap Kb = (H \cap K)c$ for some $c \in G$.
10. (Poincaré) Let G be a group and H and K be subgroups of G of finite indices. Show that $H \cap K$ is of finite index.
11. Give an example of a group G and a subgroup H of G such that $aH = bH$, but $Ha \neq Hb$ for some $a, b \in G$.
12. Let G be a group of order pq, where p and q are prime integers. Show that every proper subgroup of G is cyclic.
13. Let H be a subgroup of a group G. Define a relation $\sim$ on G by for all a, $b \in G$, $a \sim b$ if and only if $b^{-1}a \in H$ (i.e., if and only if $aH = bH$). Show that $\sim$ is an equivalence relation on G and the equivalence classes of $\sim$ are the cosets aH, $a \in G$.
14. Let $n > 1$. Show that there exists a proper subgroup H of S_n such that $[S_n : H] \leq n$.

15. Let H and K be subgroups of a finite group G such that $|H| > \sqrt{|G|}$ and $|K| > \sqrt{|G|}$. Show that $|H \cap K| > 1$.
16. Let $|G| = pq$, $(p > q)$, where p and q are distinct primes. Show that G has at most one subgroup of order p.
17. Let G be a group. If a subset A is a left coset of some subgroup of G, then show that A is a right coset of some subgroup of G.
18. Let G be a finite group and A and B be subgroups of G such that $A \subseteq B \subseteq G$. Prove that
$$[G : A] = [G : B][B : A].$$
19. Let G be a group such that $|G| < 200$. Suppose G has subgroups of order 25 and 35. Find the order of G.
20. Let G be a group of order 35 and A and B be subgroups of G of order 5 and 7, respectively. Show that $G = AB$.
21. Let A and B be subgroups of a group G. If $|A| = p$, a prime integer, show that either $A \cap B = \{e\}$ or $A \subseteq B$.
22. Let H and K be subgroups of a group G. Define a relation $\sim$ on G by for all $a, b \in G$, $a \sim b$ if and only if $b = hak$ for some $h \in H$ and $k \in K$.
    (a) Show that $\sim$ is an equivalence relation on G.
    (b) Let $a \in G$ and $[a]$ denote the equivalence class of a in G. Show that
    $$[a] = \{hak \mid h \in H,\ k \in K\} = HaK.$$
    The set HaK is called a **double coset** of H and K in G.
    (c) If G is a finite group, prove that
    $$|HaK| = \frac{|H|\,|K|}{|H \cap aKa^{-1}|}$$
    for all $a \in G$.
23. For the following, if the statement is true, then write the proof. Otherwise justify why the statement is false.
    (a) Every left coset of a subgroup of a group is also a right coset.
    (b) The product of two left cosets of a subgroup of a group is also a left coset.
    (c) There may exist a subgroup of order 12 in a group of order 40.
    (d) Let $G = \langle a \rangle$ be a cyclic group of order 30. Then $[G : \langle a^5 \rangle] = 5$.
    (e) Every proper subgroup of a group of order p^2 (p a prime) is cyclic.
    (f) Let G be a group. If H is a subgroup of order p and K is a subgroup of order q, where p and q are distinct primes, then $|HK| = pq$.

4.4 Normal Subgroups and Quotient Groups

In the previous section, we saw that a subgroup H of a group G induced two decompositions of G, one by left cosets and another by right cosets. In other words, if H is a subgroup of a group G, then G can be written as a disjoint union of distinct left (right) cosets of H in G. These two decompositions were first recognized by Galois in 1831 in the context of permutation groups. Galois called the decomposition "proper" if the two decompositions coincide, i.e., if left cosets are the same as right cosets. We call such a subgroup normal in our present-day terminology. Normal subgroups are the subject of this section. Galois showed how the solvability of a polynomial equation by means of radicals is related to the concept of a normal subgroup of the group of permutations of the roots and the group, called the quotient group, created by the normal subgroup.

Perhaps the notion of a normal subgroup is one of the most innovative ideas in group theory. I.N. Herstein (1923–1988) remarked about normal subgroups that "It is a tribute to the genius of Galois that he recognized that those subgroups for which the left and right cosets coincide are distinguished ones. Very often in mathematics the crucial problem is to recognize and to discover what are the relevant concepts; once this is accomplished the job may be more than half done."

Later C. Jordan defined normal subgroups without using the term normal as we define it in our present-day terminology.

We shall see in this text that normal subgroups play a crucial role in obtaining structural results of groups. Let us now begin our study of normal subgroups.

Definition 4.4.1 *Let G be a group. A subgroup H of G is said to be a* ***normal*** *(or* ***invariant****) subgroup of G if $aH = Ha$ for all $a \in G$.*

From the definition of a normal subgroup, it follows that for any group G, G and $\{e\}$ are normal subgroups of G.

If H is a normal subgroup of G, this does not always mean that $ah = ha$ for all $h \in H$ and for all $a \in G$ as shown by the following example.

Example 4.4.2 *Recall Example* 4.3.2. *H is a normal subgroup of S_3. Consider $h = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} \in H$. Then*

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

and

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

Hence,

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} \circ h \neq h \circ \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix},$$

even though

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

The following theorem gives a necessary and sufficient condition for a subgroup to be a normal subgroup. For $a \in G$, $\emptyset \neq H \subseteq G$, let $aHa^{-1} = \{aha^{-1} \mid h \in H\}$.

Theorem 4.4.3 *Let H be a subgroup of a group G. Then H is a normal subgroup of G if and only if for all $a \in G$, $aHa^{-1} \subseteq H$.*

Proof. First suppose that H is a normal subgroup of G. Let $a \in G$. We now show that $aHa^{-1} \subseteq H$. Let $aha^{-1} \in aHa^{-1}$, where $h \in H$. Since H is a normal subgroup of G, $aH = Ha$. Also, since $ah \in aH$, we have $ah \in Ha$ and so $ah = h'a$ for some $h' \in H$. Thus, $aha^{-1} = h' \in H$. Hence, $aHa^{-1} \subseteq H$.

Conversely, suppose $aHa^{-1} \subseteq H$ for all $a \in G$. Let $a \in G$. We show that $aH = Ha$. Let $ah \in aH$, where $h \in H$. Now $aha^{-1} \in aHa^{-1}$ and so $aha^{-1} \in H$. Thus, $aha^{-1} = h'$ for some $h' \in H$. This implies that $ah = h'a \in Ha$. Therefore, $aH \subseteq Ha$. Similarly, we can show that $Ha \subseteq aH$. Hence, $aH = Ha$. Consequently, H is a normal subgroup of G. ■

There are several other criteria that can be used to test the normality of a subgroup. We consider some of these criteria in exercises at the end of this section.

The following theorem describes some important properties of normal subgroups.

Theorem 4.4.4 *Let H and K be normal subgroups of a group G. Then*
(i) $H \cap K$ is a normal subgroup of G,
(ii) $HK = KH$ is a normal subgroup of G,
(iii) $\langle H \cup K \rangle = HK$.

Proof. (i) Since the intersection of subgroups is a subgroup, $H \cap K$ is a subgroup of G. Let $g \in G$. Consider $g(H \cap K)g^{-1}$. Let gag^{-1} be any element of $g(H \cap K)g^{-1}$, where $a \in H \cap K$. Since $a \in H \cap K$, we have $a \in H$ and $a \in K$. Hence, $gag^{-1} \in H$ and $gag^{-1} \in K$. Thus, $gag^{-1} \in H \cap K$. This shows that $g(H \cap K)g^{-1} \subseteq H \cap K$. Hence, $H \cap K$ is a normal subgroup by Theorem 4.4.3.

(ii) First we show that $HK = KH$. Let $hk \in HK$, where $h \in H$ and $k \in K$. Since K is a normal subgroup of G and $h \in G$, we have $hK = Kh$. Thus, $hk \in hK = Kh$. Since $Kh \subseteq KH$, we have $hk \in KH$. Hence, $HK \subseteq KH$. Similarly, $KH \subseteq HK$ and so $HK = KH$. Since H and K are subgroups and $HK = KH$, HK is a subgroup of G by Theorem 4.1.18. To show that HK is a normal subgroup, let $g \in G$. Then $gHg^{-1} \subseteq H$ and $gKg^{-1} \subseteq K$ since H and K are normal subgroups. Now

$$\begin{aligned} g(HK)g^{-1} &= g(Hg^{-1}gK)g^{-1} \\ &= (gHg^{-1})(gKg^{-1}) \\ &\subseteq HK. \end{aligned}$$

Therefore, HK is a normal subgroup of G by Theorem 4.4.3.

(iii) By (ii), HK is a subgroup of G. Hence, by Theorem 4.1.20,

$$HK = \langle H \cup K \rangle .$$

■

We know that if H and K are subgroups of a group G, then HK need not be a subgroup of G (Example 4.1.17). By the above theorem, if H and K are normal subgroups, then HK is a normal subgroup and hence a subgroup. However, in order to show that HK is a subgroup, we only need either H or K to be a normal subgroup. We consider one of these situations in Exercise 14 (page 93).

We now focus our attention on the study of quotient groups. First, let us consider the following example.

Example 4.4.5 *Consider the subgroup H' of Example* 4.3.2. *Now H' is not a normal subgroup of S_3. Let S_3/H' be the set of all left cosets of H' in S_3. Now let us try to define a binary operation $*$ on S_3/H'. The natural way would be to define $(\pi_1 H') * (\pi_2 H')$ to be $(\pi_1 \circ \pi_2)H'$. Now*

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

and

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

However,

$$\left(\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} H'\right) * \left(\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} H'\right) = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} H'$$

and

$$\left(\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} H'\right) * \left(\begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} H'\right) = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} H'.$$

Since

$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} H' \neq \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} H',$$

$*$ is not well defined. That $*$ is not well defined is due to the fact that H' is not a normal subgroup of S_3.

Theorem 4.4.6 *Let H be a normal subgroup of a group G. Denote the set of all left cosets $\{aH \mid a \in G\}$ by G/H and define $*$ on G/H by for all aH, $bH \in G/H$,*

$$(aH) * (bH) = abH.$$

Then $(G/H, *)$ is a group.

Proof. First we show that $*$ is well defined. Let aH, bH, $a'H$, $b'H \in G/H$ and suppose $(aH, bH) = (a'H, b'H)$. Then $aH = a'H$ and $bH = b'H$. We need to show that $aH * bH = a'H * b'H$ or $abH = a'b'H$. Now $aH = a'H$ and $bH = b'H$ imply that $a = a'h_1$ and $b = b'h_2$ for some $h_1, h_2 \in H$. Thus,

$$\begin{aligned} (a'b')^{-1}(ab) &= b'^{-1}a'^{-1}ab \\ &= b'^{-1}a'^{-1}a'h_1b'h_2 \\ &= b'^{-1}h_1b'h_2. \end{aligned}$$

Since H is a normal subgroup and $h_1 \in H$, we have $b'^{-1}h_1b'h_2 = (b'^{-1}h_1b')h_2 \in H$ and so $(a'b')^{-1}(ab) \in H$. Hence, $abH = a'b'H$ by Theorem 4.3.3(i). Thus, $*$ is well defined and so $(G/H, *)$ is a mathematical system.

Next, we show that $*$ is associative. Let aH, bH, $cH \in G/H$. Now $(aH) * [(bH) * (cH)] = (aH) * (bcH) = a(bc)H = (ab)cH = (abH) * (cH) = [(aH) * (bH)] * (cH)$. Hence, $*$ is associative. Now $eH \in G/H$ and

$$(aH) * (eH) = aeH = aH = eaH = (eH) * (aH)$$

for all $aH \in G/H$. Therefore, eH is the identity of G/H. Also, for all $aH \in G/H$, $a^{-1}H \in G/H$ and

$$(aH) * (a^{-1}H) = aa^{-1}H = eH = a^{-1}aH = (a^{-1}H) * (aH).$$

Thus, for all $aH \in G/H$, $a^{-1}H$ is the inverse of aH. Consequently, $(G/H, *)$ is a group. ■

Definition 4.4.7 *Let G be a group and H be a normal subgroup of G. The group G/H is called the* ***quotient group*** *of G by H.*

Example 4.4.8 *Consider the subgroup* $(\langle n\rangle, +)$ *of the group* $(\mathbb{Z}, +)$*, where* n *is a fixed positive integer. Since* $\mathbb{Z}$ *is commutative,* $\langle n\rangle$ *is a normal subgroup of* $\mathbb{Z}$ *(Exercise* 16*, page* 93*). Hence,* $(\mathbb{Z}/\langle n\rangle, +)$ *is a group, where*

$$(a + \langle n\rangle) + (b + \langle n\rangle) = (a + b) + \langle n\rangle$$

for all $a + \langle n\rangle, b + \langle n\rangle \in \mathbb{Z}/\langle n\rangle$*. In Example* 4.3.10*, we determined the distinct left cosets of* $\langle n\rangle$ *in* $\mathbb{Z}$*. We found that*

$$\mathbb{Z}/\langle n\rangle = \{0 + \langle n\rangle, 1 + \langle n\rangle, 2 + \langle n\rangle, \ldots, n - 1 + \langle n\rangle\}.$$

Example 4.4.9 *Consider the normal subgroup* H *of* S_3 *of Example* 4.4.2*. Since* $|S_3| = 6$ *and* $|H| = 3$*,* $[S_3 : H] = 2$ *by Lagrange's theorem. Now* $|S_3/H| = [S_3 : H] = 2$ *and for all* $h \in H$*,* $hH = H$*. Thus,* $eH = H$*,* $(1\ 2\ 3)H = H$ *and* $(1\ 3\ 2)H = H$*. We have shown in Example* 4.3.2 *that* $(2\ 3)H = (1\ 3)H = (1\ 2)H$*. Thus,*

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

We also note that S_3/H *is cyclic and* $(2\ 3)H$ *is a generator for* S_3/H*.*

Example 4.4.10 *Consider* $\mathbb{Z}_8$ *and let* $H = \{[0], [4]\}$*. Then* H *is a normal subgroup of* $\mathbb{Z}_8$*. Now* $|H| = 2$ *and* $|\mathbb{Z}_8| = 8$*. Thus,* $|\mathbb{Z}_8/H| = \frac{|\mathbb{Z}_8|}{|H|} = 4$*. Hence,* $\mathbb{Z}_8/H$ *has four elements. Now*

$$[0] + H = H = [4] + H,$$

$$[1] + H = \{[1], [5]\} = [5] + H,$$

$$[2] + H = \{[2], [6]\} = [6] + H,$$

and

$$[3] + H = \{[3], [7]\} = [7] + H.$$

Hence, $\mathbb{Z}_8/H = \{[0] + H, [1] + H, [2] + H, [3] + H\}$*.*

Example 4.4.11 *Consider* $\mathbb{Z}_4 \times \mathbb{Z}_6$*, the direct product of* $\mathbb{Z}_4$ *and* $\mathbb{Z}_6$*. Let*

$$H = \langle([0], [1])\rangle = \{([0], [0]), ([0], [1]), ([0], [2]), ([0], [3]), ([0], [4]), ([0], [5])\}.$$

Then H *is a subgroup of* $\mathbb{Z}_4 \times \mathbb{Z}_6$ *and since* $\mathbb{Z}_4 \times \mathbb{Z}_6$ *is commutative,* H *is a normal subgroup of* $\mathbb{Z}_4 \times \mathbb{Z}_6$*. Now* $|\mathbb{Z}_4 \times \mathbb{Z}_6| = 24$ *and* $|H| = 6$*. Hence,*

$$|(\mathbb{Z}_4 \times \mathbb{Z}_6)/H| = \frac{|\mathbb{Z}_4 \times \mathbb{Z}_6|}{|H|} = 4.$$

Thus, $(\mathbb{Z}_4 \times \mathbb{Z}_6)/H$ *has four elements. Since for all* $[n] \in \mathbb{Z}_6$*,* $([0], [n]) \in H$*, we have for all* $[n] \in \mathbb{Z}_6$*,* $([0], [n]) + H = H$*. Let* $([m], [n]) \in \mathbb{Z}_4 \times \mathbb{Z}_6$*. Then* $([m], [n]) = ([m], [0]) + ([0], [n])$ *and from this, it follows that* $([m], [n]) + H = ([m], [0]) + H$*. Let us now compute* $([m], [0]) + H$ *for* $m = 0, 1, 2, 3$*. Now* $([0], [0]) + H = H$*,*

$$([1], [0]) + H = \{([1], [0]), ([1], [1]), ([1], [2]), ([1], [3]), ([1], [4]), ([1], [5])\},$$

$$([2], [0]) + H = \{([2], [0]), ([2], [1]), ([2], [2]), ([2], [3]), ([2], [4]), ([2], [5])\},$$

and

$$([3], [0]) + H = \{([3], [0]), ([3], [1]), ([3], [2]), ([3], [3]), ([3], [4]), ([3], [5])\}.$$

From above, we see that $([0], [0]) + H$*,* $([1], [0]) + H$*,* $([2], [0]) + H$*, and* $([3], [0]) + H$ *are all distinct. Hence,*

$$(\mathbb{Z}_4 \times \mathbb{Z}_6)/H = \{([0], [0]) + H, ([1], [0]) + H, ([2], [0]) + H, ([3], [0]) + H\}.$$

Worked-Out Exercises

♢ **Exercise 1** Let H be a subgroup of a group G. Then $W = \cap_{g\in G} gHg^{-1}$ is a normal subgroup of G.

Solution: By Worked-Out Exercise 1 (page 75), gHg^{-1} is a subgroup of G for all $g \in G$. Since the intersection of subgroups is a subgroup, W is a subgroup of G. Let $x \in G$, $w \in W$. Then $w \in gHg^{-1}$ for all $g \in G$. We show that $xwx^{-1} \in gHg^{-1}$ for all $g \in G$, which in turn will yield that $xwx^{-1} \in W$. Let $g \in G$.

Let us work our way backward and suppose $xwx^{-1} \in gHg^{-1}$. Then $xwx^{-1} = ghg^{-1}$ for some $h \in H$. Thus, $g^{-1}xw\ x^{-1}g = h \in H$. This implies that

$$(g^{-1}x)w(g^{-1}x)^{-1} \in H.$$

Set $y = x^{-1}g$. Then $g = xy$. Hence, in order to show that $xwx^{-1} \in gHg^{-1}$ for a given $g \in G$, first we need to find $y \in G$ such that $g = xy$. Since $g = x(x^{-1}g)$, we can choose $y = x^{-1}g$.

So there exists $y \in G$ such that $g = xy$. Since $y \in G$, we have $w \in yHy^{-1}$ and so $w = yhy^{-1}$ for some $h \in H$. Therefore, $xwx^{-1} = x(yhy^{-1})x^{-1} = xyhy^{-1}x^{-1} = (xy)h(xy)^{-1} = ghg^{-1} \in gHg^{-1}$. Since $g \in G$ was arbitrary, $xwx^{-1} \in gHg^{-1}$ for all $g \in G$. Consequently, W is a normal subgroup of G.

♢ **Exercise 2** Let H be a subgroup of G.

(a) If $x^2 \in H$ for all $x \in G$, prove that H is a normal subgroup of G and G/H is commutative.

(b) If $[G : H] = 2$, prove that H is a normal subgroup of G.

Solution: (a) Let $g \in G$ and $h \in H$. Consider ghg^{-1} and note that

$$ghg^{-1} = (gh)^2h^{-1}g^{-2}.$$

Now $h^{-1} \in H$ and by our hypothesis $(gh)^2$, $g^{-2} \in H$. This implies that $ghg^{-1} \in H$, which in turn shows that $gHg^{-1} \subseteq H$. Hence, H is a normal subgroup of G. To show that G/H is commutative, let $xH, yH \in G/H$. We wish to show that $xHyH = yHxH$ or $xyH = yxH$ or $(yx)^{-1}(xy) \in H$. Consider $(yx)^{-1}(xy)$. Now

$$(yx)^{-1}(xy) = (x^{-1}y^{-1})(xy) = (x^{-1}y^{-1})^2(yxy^{-1})^2y^2.$$

Since $a^2 \in H$ for all $a \in G$, it follows that $(x^{-1}y^{-1})^2(yxy^{-1})^2y^2 \in H$ and so $(yx)^{-1}(xy) \in H$. Thus, G/H is commutative.

(b) We prove that H is a normal subgroup of G first by showing that $x^2 \in H$ for all $x \in G$ and then by using (i). Suppose there exists $x \in G$ such that $x^2 \notin H$. Then $x \notin H$ and so H and xH are distinct left cosets of H in G. Since $[G : H] = 2$, it follows that $G/H = \{H, xH\}$. Hence, $G = H \cup xH$. This implies that $x^2 \in H \cup xH$. Since $x^2 \notin H$, we must have $x^2 \in xH$. Hence, $x^2 = xh$ for some $h \in H$. But then $x = h \in H$, which is a contradiction. Hence, $x^2 \in H$ for all $x \in G$. By (i), H is a normal subgroup of G.

Exercise 3 Let G be a group such that every cyclic subgroup of G is a normal subgroup of G. Prove that every subgroup of G is a normal subgroup of G.

Solution: Let H be a subgroup of G. Let $g \in G$ and $a \in H$. Then $g^{-1}ag \in \langle a\rangle \subseteq H$. Hence, H is normal in G.

♢ **Exercise 4** Let H be a proper subgroup of G such that for all $x, y \in G\backslash H$, $xy \in H$. Prove that H is a normal subgroup of G.

Solution: Let $x \in G\backslash H$. Then $x^{-1} \in G\backslash H$. Let $y \in H$. Then $xy \in G\backslash H$. Thus, xy, $x^{-1} \in G\backslash H$. Hence, $xyx^{-1} \in H$. Therefore, H is a normal subgroup of G.

♢ **Exercise 5** Let G be a group and $\{N_i \mid i \in \Omega\}$ be a family of proper normal subgroups of G. Suppose $G = \cup_i N_i$ and $N_i \cap N_j = \{e\}$ for $i \neq j$. Prove that G is commutative.

Solution: Let $x, y \in G$. Then there exist i and j such that $x \in N_i$ and $y \in N_j$. If $i \neq j$, then since $N_i \cap N_j = \{e\}$, $xy = yx$ (Exercise 13, page 93). Let $i = j$. Now there exists $z \in G$ such that $z \notin N_i$. Then $zx \notin N_i$. Hence, $zx \in N_l$ for some $l \neq i$ and so $(zx)y = y(zx)$. Thus, $z(xy) = (zx)y = y(zx) = (yz)x = (zy)x = z(yx)$. This implies that $xy = yx$. Consequently, G is commutative.

Exercise 6 Let H be a subgroup of a group G. Suppose that the product of two left cosets of H in G is again a left coset of H in G. Prove that H is a normal subgroup of G.

Solution: Let $g \in G$. Then $gHg^{-1}H = tH$ for some $t \in G$. Thus, $e = geg^{-1}e \in tH$. Hence, $e = th$ for some $h \in H$. Thus, $t = h^{-1} \in H$ so that $tH = H$. Now $gHg^{-1} \subseteq gHg^{-1}H = H$. Therefore, H is a normal subgroup of G.

♢ **Exercise 7** Let G be a group. Show that if $G/Z(G)$ is cyclic, then G is commutative.

Solution: Write $Z = Z(G)$. Let $G/Z = \langle gZ \rangle$. Let $a, b \in G$. Then $aZ, bZ \in G/Z$. Hence, $aZ = g^n Z$ and $bZ = g^m Z$ for some $n, m \in \mathbb{Z}$. Then $a \in g^n Z$ and $b \in g^m Z$. Thus, $a = g^n d$ and $b = g^m h$ for some $d, h \in Z$. Now $ab = g^n d g^m h = g^n g^m dh$ (since $d \in Z$) $= g^{n+m} hd$ (since $h \in Z$) $= g^m g^n hd = g^m h g^n d = ba$. Hence, G is commutative.

Exercises

1. Let
$$H = \left\{ e, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} \right\},$$
where e is the identity permutation. Determine whether or not H is a normal subgroup of S_4.
2. Let H denote the subgroup $\{r_{360}, h\}$ of the group of symmetries of the square. Determine whether or not H is a normal subgroup of G.
3. Let G be a group and H be a subgroup of G. Show that H is normal if and only if $ghg^{-1} \in H$ for all $g \in G$, $h \in H$.
4. Let G be a group and H be a subgroup of G. If for all $a, b \in G$, $ab \in H$ implies $ba \in H$, prove that H is a normal subgroup of G.
5. Let H be a proper subgroup of a group G and $a \in G$, $a \notin H$. Suppose that for all $b \in G$, either $b \in H$ or $Ha = Hb$. Show that H is a normal subgroup of G.
6. Let G be a group. Prove that $Z(G)$ is a normal subgroup of G.
7. Let G be a group. Let H be a subgroup of G such that $H \subseteq Z(G)$. Show that if G/H is cyclic, then $G = Z(G)$, i.e., G is commutative.
8. Let H and K be subgroups of a group G such that H is a normal subgroup of G. Prove that $H \cap K$ is a normal subgroup of K.
9. Determine the quotient groups of
   (a) $(\mathbb{E}, +)$ in $(\mathbb{Z}, +)$,
   (b) $(\mathbb{Z}, +)$ in $(\mathbb{Q}, +)$,
   (c) $(\langle [4] \rangle, +_{12})$ in $(\mathbb{Z}_{12}, +_{12})$
   (d) $(4\mathbb{Z}, +)$ in $(\mathbb{Z}, +)$.
10. Let H be a normal subgroup of a group G. Prove that if G is commutative, then so is the quotient group G/H.
11. If H is a subgroup of finite index in $(\mathbb{C}^*, \cdot)$,then prove that $H = \mathbb{C}^*$
12. Let H be a nonempty subset of a group G. The set $N(H) = \{a \in G \mid aHa^{-1} = H\}$ is called the **normalizer** of H in G.
   (a) Prove that $N(H)$ is a subgroup of G.
   Suppose H is a subgroup of G.
   (b) Prove that H is normal in G if and only if $N(H) = G$.
   (c) Prove that H is normal in $N(H)$.
   (d) Prove that $N(H)$ is the largest subgroup of G in which H is normal, i.e., if H is normal in a subgroup K of G, then $K \subseteq N(H)$.
13. Let H and K be normal subgroups of a group G. If $H \cap K = \{e\}$, prove that $hk = kh$ for all $h \in H$ and $k \in K$.
14. Let G be a group. Let H be a subgroup of G and K be a normal subgroup of G. Prove that HK is a subgroup of G.
15. Give an example of a noncommutative group in which every subgroup is normal.
16. Show that every subgroup of a commutative group is normal.
17. Let H be a normal subgroup of a group G such that $|H| = 2$. Show that $H \subseteq Z(G)$.
18. Show that if H is the only subgroup of order n in a group G, then H is a normal subgroup of G.

19. Let $K = \{e,\ (1\ 2) \circ (3\ 4),\ (1\ 4) \circ (3\ 2),\ (1\ 3) \circ (2\ 4)\}$.

    (a) Show that K is the only subgroup of order 4 in A_4.

    (b) Show that K is a normal subgroup of A_4.

20. Show that A_4 has no subgroup of order 6.

21. Find all subgroups of A_4.

22. Prove that A_n is the only subgroup of index 2 in S_n.

23. Let G be a group. An equivalence relation ρ on G is called a **congruence relation** if

$$\text{for all } a, b, c \in G, a\rho b \text{ implies that } ca\rho cb \text{ and } ac\rho bc.$$

    Let H be a normal subgroup of G. Define the relation ρ_H on G by

$$\text{for all } a, b \in G, a\rho_H b \text{ if and only if } a^{-1}b \in H.$$

    Prove that

    (a) ρ_H is a congruence relation on G,

    (b) The ρ_H class $a\rho_H = \{b \in G \mid a\rho_H b\}$ is the left coset aH,

    (c) $H = e\rho_H$.

24. Let H be a subgroup of a group G. Define a relation ρ_H on G by $\rho_H = \{(a, b) \in G \times G \mid a^{-1}b \in H\}$. Show that if ρ_H is a congruence relation, then H is a normal subgroup of G.

25. Let ρ be a congruence relation on a group G. Show that there exists a normal subgroup H of G such that $\rho = \{(a, b) \in G \times G \mid a^{-1}b \in H\}$.

26. For the following statements, write the proof if the statement is true; otherwise, give a counterexample.

    (i) A subgroup H of a group G is a normal subgroup if and only if every right coset of H is also a left coset.

    (ii) If A, B and C are normal subgroups of a group G, then $A(B \cap C)$ is a normal subgroup of G.

    (iii) If A is a normal subgroup of a finite group G, then $[G : A] = 2$.

    (iv) Every commutative subgroup of a group G is a normal subgroup of G.

    (v) If G is a group of order $2p$, p an odd prime, then either G is commutative or G contains a normal subgroup of order p.

    (vi) If every element of a group G is of finite order, then G is a finite group.

    (vii) A_5 is the only nontrivial normal subgroup S_5.

Joseph Louis Lagrange (1736–1813) was born on January 25, 1736, in Turin, Italy. He spent the early part of his life in Turin. While there he was involved in carrying out research work in calculus of variations and mechanics.

In 1766, Lagrange was invited by the Prussian king, Frederick II, to fill the position vacated by Euler in Berlin. Frederick the Great proclaimed in his appointment that "the greatest king in Europe" ought to have "the greatest mathematician in Europe." In 1787, after the death of Frederick II, he went to Paris, accepting an invitation from Louis XVI. In 1797, he accepted a position at the newly formed École Polytechnique in Paris. He was made a count by Napoleon and remained at the École Polytechnique till his death. He died on April 10, 1813.

Throughout his life, Lagrange did work of fundamental importance. He made numerous contributions to many branches of mathematics, including number theory, the theory of equations, differential equations, celestial mechanics, and fluid mechanics. In 1770, he proved the famous Lagrange's theorem in group theory.

He is responsible for the work leading to Galois theory. In his paper, "Réflexion sur la théorie algébriques des équations," Lagrange carefully analyzed the various known methods to solve a polynomial equation of degree ≤ 4 by means of radicals. He was interested in finding a general method of solution for polynomials of higher degree. He was unable to find a general solution, but in his paper he introduced several key ideas on the permutations of roots which finally led Abel and Galois to develop the necessary theory to answer the question. Lagrange's work on the solution of polynomial equations is one of the sources from which modern group theory evolved.

Chapter 5

Homomorphisms and Isomorphisms of Groups

One of the main uses of the concept of an isomorphism is the classification of algebraic structures−in particular, groups. Readers with some knowledge of linear algebra may recall that the concept of an isomorphism is used to completely characterize vector spaces with the same field of scalars in terms of a single integer, the dimension of the vector space. Another important use of an isomorphism is the representation of one algebraic structure by means of another. This is done in linear algebra, where it is shown that the vector space of all linear transformations from one finite dimensional vector space into another is isomorphic to a certain vector space of matrices.

5.1 Homomorphisms of Groups

In this section, we consider certain mappings between groups. These mappings will be defined in such a way as to preserve the algebraic structure of the groups involved. More precisely, suppose we are given a function f from a group G into a group G_1, where $*_1$ denotes the operation of G_1. Let $a, b \in G$. Then under f, a corresponds to $f(a)$, b to $f(b)$, and $a * b$ to $f(a * b)$. If f is to preserve the operations of G and G_1, $a * b$ must correspond to $f(a) *_1 f(b)$. Since f is a function, this forces the requirement that $f(a * b) = f(a) *_1 f(b)$.

Definition 5.1.1 *Let $(G, *)$ and $(G_1, *_1)$ be groups and f a function from G into G_1. Then f is called a **homomorphism** of G into G_1 if for all $a, b \in G$,*

$$f(a * b) = f(a) *_1 f(b).$$

Let the identity element of the group G_1 be denoted by e_1.

Define $f : G \to G_1$ by $f(a) = e_1$ for all $a \in G$. Since $f(a * b) = e_1 = e_1 *_1 e_1 = f(a) *_1 f(b)$ for all $a, b \in G$, we find that f is a homomorphism from G into G_1. This shows that there always exists a homomorphism from a group G into a group G_1. This homomorphism is called the **trivial homomorphism**.

The identity map from G onto G is also a homomorphism.

Before we consider more examples of homomorphisms, let us prove some basic properties of homomorphisms.

Theorem 5.1.2 *Let f be a homomorphism of a group G into a group G_1. Then*

(i) $f(e) = e_1$.

(ii) $f(a^{-1}) = f(a)^{-1}$ for all $a \in G$.

(iii) If H is a subgroup of G, then $f(H) = \{f(h) \mid h \in H\}$ is a subgroup of G_1.

(iv) If H_1 is a subgroup of G_1, then $f^{-1}(H_1) = \{g \in G \mid f(g) \in H_1\}$ is a subgroup of G, and if H_1 is a normal subgroup, then $f^{-1}(H_1)$ is a normal subgroup of G.

(v) If G is commutative, then $f(G)$ is commutative.

(vi) If $a \in G$ is such that $\circ(a) = n$, then $\circ(f(a))$ divides n.

Proof. (i) Since f is a homomorphism, $f(e)f(e) = f(ee) = f(e) = f(e)e_1$. This implies that $f(e) = e_1$ by the cancellation law.

(ii) Let $a \in G$. Then $f(a)f(a^{-1}) = f(aa^{-1}) = f(e) = e_1$. Similarly, $f(a^{-1})f(a) = e_1$. Since $f(a)$ has a unique inverse, $f(a^{-1}) = f(a)^{-1}$.

(iii) Let H be a subgroup of G. Then $e \in H$ and by (i), $f(e) = e_1$. Thus, $e_1 = f(e) \in f(H)$ and so $f(H) \neq \emptyset$. Let $f(a), f(b) \in f(H)$, where $a, b \in H$. Since H is a subgroup, $ab^{-1} \in H$. Thus, $f(a)f(b)^{-1} = f(a)f(b^{-1}) = f(ab^{-1}) \in f(H)$. Hence, by Theorem 4.1.6, $f(H)$ is a subgroup of G_1.

(iv) By (i), $e \in f^{-1}(H_1)$ and so $f^{-1}(H_1) \neq \emptyset$. Let $a, b \in f^{-1}(H_1)$. Then $f(a)$, $f(b) \in H_1$. Hence, $f(ab^{-1}) = f(a)f(b^{-1}) = f(a)f(b)^{-1} \in H_1$ and so $ab^{-1} \in f^{-1}(H_1)$. Thus, by Theorem 4.1.6, $f^{-1}(H_1)$ is a subgroup of G. Suppose H_1 is a normal subgroup of G_1. Let $g \in G$. We now show that $gf^{-1}(H_1)g^{-1} \subseteq f^{-1}(H_1)$. Let $a \in gf^{-1}(H_1)g^{-1}$. Then $a = gbg^{-1}$ for some $b \in f^{-1}(H_1)$. Now $f(a) = f(gbg^{-1}) = f(g)f(b)f(g^{-1}) = f(g)f(b)f(g)^{-1} \in H_1$ since H_1 is a normal subgroup of G_1 and $f(b) \in H_1$. Hence, $a \in f^{-1}(H_1)$ and this shows that $gf^{-1}(H_1)g^{-1} \subseteq f^{-1}(H_1)$. Thus, $f^{-1}(H_1)$ is a normal subgroup of G.

(v) Suppose G is commutative. Let $f(a)$, $f(b) \in f(G)$. Then $f(a)f(b) = f(ab) = f(ba) = f(b)f(a)$. Hence, $f(G)$ is commutative.

(vi) Since $(f(a))^n = f(a^n) = f(e) = e_1$, we have $\circ(f(a))$ divides n by Theorem 2.1.46. ■

Definition 5.1.3 *Let f be a homomorphism of a group G into a group G_1. The* ***kernel*** *of f, written Ker f, is defined to be the set*

$$Ker\ f = \{a \in G \mid f(a) = e_1\}.$$

By Theorem 5.1.2, $e \in$ Ker f.

Example 5.1.4 *Define the function f from $(\mathbb{Z}, +)$ into $(\mathbb{Z}_n, +_n)$ by $f(a) = [a]$ for all $a \in \mathbb{Z}$. From the definition of f, it follows that f maps $\mathbb{Z}$ onto $\mathbb{Z}_n$. Let $a, b \in \mathbb{Z}$. Then*

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

Thus, f is a homomorphism of $\mathbb{Z}$ onto $\mathbb{Z}_n$. Now

$$\begin{aligned} Ker\ f &= \{a \in \mathbb{Z} \mid f(a) = [0]\} \\ &= \{a \in \mathbb{Z} \mid [a] = [0]\} \\ &= \{a \in \mathbb{Z} \mid a \text{ is divisible by } n\} \\ &= \{a \in \mathbb{Z} \mid a = qn \text{ for some } q \in \mathbb{Z}\} \\ &= \{qn \mid q \in \mathbb{Z}\}. \end{aligned}$$

The above example shows that a nontrivial finite group may be an image of an infinite group under a homomorphism. By Theorem 5.1.2(v), a noncommutative group cannot be an image under a homomorphism of a commutative group. In the next example, we show that two finite groups G and G_1 having same number of elements need not have a homomorphism from G onto G_1.

Example 5.1.5 *The groups $\mathbb{Z}_4 \times \mathbb{Z}_4$ and $\mathbb{Z}_8 \times \mathbb{Z}_2$ are commutative and each is of order* 16. *Suppose there exists a homomorphism f of $\mathbb{Z}_4 \times \mathbb{Z}_4$ onto $\mathbb{Z}_8 \times \mathbb{Z}_2$. Now $a = ([7], [0]) \in \mathbb{Z}_8 \times \mathbb{Z}_2$ and $\circ(a) = 8$. Since f is onto $\mathbb{Z}_8 \times \mathbb{Z}_2$, there exists $b \in \mathbb{Z}_4 \times \mathbb{Z}_4$ such that $f(b) = a$. By Theorem* 5.1.2(*vi*), $\circ(f(b))$ *divides* $\circ(b)$. *Since* $\circ(f(b)) = 8$ *and* $\mathbb{Z}_4 \times \mathbb{Z}_4$ *has elements of order* 1, 2, *and* 4 *only,* $\circ(f(b))$ *cannot divide* $\circ(b)$. *This is a contradiction. Hence, there does not exist any homomorphism from $\mathbb{Z}_4 \times \mathbb{Z}_4$ onto $\mathbb{Z}_8 \times \mathbb{Z}_2$.*

Definition 5.1.6 *Let G and G_1 be groups. A homomorphism $f : G \to G_1$ is called an* ***epimorphism*** *if f is onto G_1 and f is called a* ***monomorphism*** *if f is one-one. If there is an epimorphism f from G onto G_1, then G_1 is called a* ***homomorphic image*** *of G.*

The homomorphism in Example 5.1.4 is an epimorphism, but not a monomorphism.

Example 5.1.7 *Let $\mathbb{R}^*$ be the group of all nonzero real numbers under multiplication. Define $f : \mathbb{R}^* \to \mathbb{R}^*$ by $f(a) = |a|$. Now $f(ab) = |ab| = |a|\,|b| = f(a)f(b)$, which implies that f is a homomorphism. Since $f(1) = 1 = f(-1)$ and $1 \neq -1$, f is not one-one. Also, from the definition of f, it follows that f is not onto $\mathbb{R}^*$. Hence, f is neither an epimorphism nor a monomorphism.*

The following theorem gives a necessary and sufficient condition for a homomorphism to be a one-one mapping in terms of its kernel.

Theorem 5.1.8 *Let f be a homomorphism of a group G into a group G_1. Then f is one-one if and only if Ker $f = \{e\}$.*

Proof. Suppose f is one-one. Let $a \in \text{Ker } f$. Then $f(a) = e_1 = f(e)$ by Theorem 5.1.2(i). Since f is one-one, we must have $a = e$. Hence, Ker $f = \{e\}$. Conversely, suppose that Ker $f = \{e\}$. Let $a, b \in G$. Suppose $f(a) = f(b)$. Then

$$f(ab^{-1}) = f(a)f(b^{-1}) = f(a)f(b)^{-1} = e_1.$$

Thus, $ab^{-1} \in \text{Ker } f = \{e\}$ and so $ab^{-1} = e$, i.e., $a = b$. This proves that f is one-one. ■

Theorem 5.1.9 *Let f be a homomorphism of a group G into a group G_1. Then Ker f is a normal subgroup of G.*

Proof. Since $e \in \text{Ker } f$, Ker $f \neq \emptyset$. Let $a, b \in \text{Ker } f$. Then $f(ab^{-1}) = f(a)f(b^{-1}) = f(a)f(b)^{-1} = e_1(e_1)^{-1} = e_1e_1 = e_1$. Thus, $ab^{-1} \in \text{Ker } f$ and hence Ker f is a subgroup of G by Theorem 4.1.6. Let $a \in G$ and $h \in \text{Ker } f$. Then $f(aha^{-1}) = f(a)f(h)f(a^{-1}) = f(a)f(h)f(a)^{-1} = f(a)e_1f(a)^{-1} = e_1$. Therefore, $aha^{-1} \in \text{Ker } f$. This proves that $a\text{Ker } fa^{-1} \subseteq \text{Ker } f$. Hence, Ker f is a normal subgroup of G by Theorem 4.4.3. ■

Example 5.1.10 *Let* $GL(2,\mathbb{R}) = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a, b, c, d \in \mathbb{R},\ ad - bc \neq 0 \right\}$ *be the noncommutative group of Example* 2.1.14. *Let $\mathbb{R}^*$ be the group of all nonzero real numbers under multiplication. Define $f : GL(2,\mathbb{R}) \to \mathbb{R}^*$ by*

$$f\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) = ad - bc$$

for all $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in GL(2,\mathbb{R})$. *Let* $\begin{bmatrix} a & b \\ c & d \end{bmatrix}, \begin{bmatrix} u & v \\ w & s \end{bmatrix} \in GL(2,\mathbb{R})$. *Now*

$$\begin{aligned} f\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} u & v \\ w & s \end{bmatrix} \right) &= f\left(\begin{bmatrix} au + bw & av + bs \\ cu + dw & cv + ds \end{bmatrix} \right) \\ &= (au + bw)(cv + ds) - (av + bs)(cu + dw) \\ &= (ad - bc)(us - vw) \\ &= f\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) f\left(\begin{bmatrix} u & v \\ w & s \end{bmatrix} \right). \end{aligned}$$

This proves that f is a homomorphism. To show that f is onto $\mathbb{R}^*$, let $a \in \mathbb{R}^*$. Then $\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} \in GL(2,\mathbb{R})$ *and* $f\left(\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} \right) = a$. *Hence, f is onto $\mathbb{R}^*$. Since* $f\left(\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} \right) = a = f\left(\begin{bmatrix} a & 1 \\ 0 & 1 \end{bmatrix} \right)$ *and* $\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} \neq \begin{bmatrix} a & 1 \\ 0 & 1 \end{bmatrix}$, *$f$ is not one-one.*

The previous example shows that there may exist a homomorphism of a noncommutative group onto a commutative group.

Example 5.1.11 *Consider S_3 and the normal subgroup*

$$H = \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \right\}.$$

Define $f : S_3 \to S_3/H$ by for all $\pi \in S_3$, $f(\pi) = \pi H$. Then

$$f(\pi \circ \pi') = (\pi \circ \pi')H = (\pi H) \circ (\pi' H) = f(\pi) \circ f(\pi')$$

for all $\pi, \pi' \in S_3$. Hence, f is a homomorphism. Also, Ker $f = \{\alpha \in S_3 \mid \alpha H = H\} = \{\alpha \in S_3 \mid \alpha \in H\} = H$.

In Theorem 5.1.9, we showed that if f is a homomorphism of a group into a group G_1, then Ker f is a normal subgroup of G. In the following theorem, we show that every normal subgroup H of a group G induces a homomorphism g of G onto the quotient group G/H such that Ker $g = H$. We note that in Example 5.1.11, the conclusion did not depend on the nature of S_3. The conclusion was made by use of general arguments. This also leads us to the following theorem.

Theorem 5.1.12 *Let H be a normal subgroup of a group G. Define the function g from G onto the quotient group G/H by $g(a) = aH$ for all $a \in G$. Then g is a homomorphism of G onto G/H and Ker $g = H$. (The homomorphism g is called the* ***natural homomorphism*** *of G onto G/H.)*

Proof. From the definition of g, it follows that g is a function from G onto G/H. To show g is a homomorphism, let $a, b \in G$. Then $g(ab) = (ab)H = (aH)(bH) = g(a)g(b)$. Hence, g is a homomorphism of G onto G/H. Finally, we show that Ker $g = H$. Now $a \in$ Ker g if and only if $g(a) = eH$ if and only if $aH = eH$ if and only if $e^{-1}a \in H$ if and only if $a \in H$. Thus, Ker $g = H$. ■

We now define a particular type of homomorphism between groups in order to introduce the important idea of groups being algebraically indistinguishable.

Definition 5.1.13 *A homomorphism f of a group G into a group G_1 is called an **isomorphism** of G onto G_1 if f is one-one and onto G_1. In this case, we write $G \simeq G_1$ and say that G and G_1 are **isomorphic**. An isomorphism of a group G onto G is called an **automorphism**.*

For a group G, Aut(G), denotes the set of all automorphisms of G.

In the following theorem, we collect some properties of isomorphisms, which will be useful in determining whether given groups are isomorphic or not.

Theorem 5.1.14 *Let f be an isomorphism of a group G onto a group G_1. Then*

(i) $f^{-1} : G_1 \to G$ is an isomorphism.

(ii) G is commutative if and only if G_1 is commutative.

(iii) For all $a \in G$, $\circ(a) = \circ(f(a))$.

(iv) G is a torsion group if and only if G_1 is a torsion group.

(v) G is cyclic if and only if G_1 is cyclic.

Proof. (i) Since f is one-one and onto G_1, f^{-1} is one-one and onto G. Now we only need to verify that f^{-1} is a homomorphism. Let $u, v \in G_1$. Then there exist $a, b \in G$ such that $f(a) = u$ and $f(b) = v$. This implies that $a = f^{-1}(u)$, $b = f^{-1}(v)$, and $uv = f(a)f(b) = f(ab)$. Thus, $f^{-1}(uv) = ab = f^{-1}(u)f^{-1}(v)$ and so f^{-1} is a homomorphism. Hence, f^{-1} is an isomorphism.

(ii) Suppose G is commutative. Let $u, v \in G_1$. Since f is onto G_1, there exist $a, b \in G$ such that $f(a) = u$ and $f(b) = v$. Now

$$uv = f(a)f(b) = f(ab) = f(ba) = f(b)f(a) = vu.$$

Thus, G_1 is commutative. Conversely, suppose G_1 is commutative. Let $a, b \in G$. Now

$$f(ab) = f(a)f(b) = f(b)f(a) = f(ba).$$

Since f is one-one, we have $ab = ba$. This proves that G is commutative.

(iii) Let $a \in G$. By induction, it follows that for all positive integers n, $f(a^n) = (f(a))^n$. Since f is one-one, for all $b \in G$, $f(b) = e_1$ if and only if $b = e$. Hence, $a^n = e$ if and only if $(f(a))^n = e_1$. Thus, a is of finite order if and only if $f(a)$ is of finite order. Suppose $\circ(a) = m$ and $\circ(f(a)) = n$. Since $a^m = e$, $(f(a))^m = e_1$. By Theorem 2.1.46, n divides m. Also, $(f(a))^n = e_1$ implies that $a^n = e$. Hence, m divides n. Since m and n are both positive integers and m divides n and n divides m, it follows that $m = n$.

(iv) This follows immediately by (iii).

(v) Suppose G is cyclic. Then $G = \langle a \rangle$ for some $a \in G$. Since $f(a) \in G_1$, $\langle f(a) \rangle \subseteq G_1$. Let $b \in G_1$. Since f is onto G_1, there exists $c \in G$ such that $f(c) = b$. Now $c = a^n$ for some $n \in \mathbb{Z}$. Thus,

$$b = f(c) = f(a^n) = (f(a))^n \in \langle f(a) \rangle .$$

Hence, $G_1 = \langle f(a) \rangle$ and so G_1 is cyclic. The converse follows since f^{-1} is an isomorphism. ■

In order to develop a feel for two groups being algebraically indistinguishable, let us consider two sets S and S' such that there is a one-one function f of S onto S'. Then in a set-theoretic sense, S and S' are the same sets "under f". For instance, let A and B be subsets of S. Then $f(A)$ and $f(B)$ are corresponding subsets of S'. Now $f(A \cap B) = f(A) \cap f(B)$ and $f(A \cup B) = f(A) \cup f(B)$; that is, union and intersection are preserved under f. Other purely set-theoretic operations can be seen to be preserved under f also. Now suppose binary operations $*$ and $*'$ are defined on S and S', respectively, so that $(S, *)$ and $(S', *')$ are groups. Now even though S and S' are the same sets "under f," they need not be the same as groups, i.e., f may not preserve operations. We have seen that the requirement for f to preserve operations is that $f(a * b) = f(a) *' f(b)$ for all $a, b \in S$.

We now consider examples of groups that are isomorphic and examples of groups that are not isomorphic.

Example 5.1.15 *Let n be a positive integer. Define f from $\mathbb{Z}_n$ into $\mathbb{Z}/\langle n \rangle$ by for all $[a] \in \mathbb{Z}_n$, $f([a]) = a + \langle n \rangle$. Then $[a] = [b]$ if and only if $n \mid (a - b)$ if and only if $a - b = nq$ for some $q \in \mathbb{Z}$ if and only if $a - b \in \langle n \rangle$ if and only if $a + \langle n \rangle = b + \langle n \rangle$ if and only if $f([a]) = f([b])$. Therefore, we find that f is a one-one function. From the definition of f, it follows that f maps $\mathbb{Z}_n$ onto $\mathbb{Z}/\langle n \rangle$. Now $f([a] +_n [b]) = f([a+b]) = (a+b) + \langle n \rangle = (a + \langle n \rangle) + (b + \langle n \rangle) = f([a]) + f([b])$. Thus, f is an isomorphism of $\mathbb{Z}_n$ onto $\mathbb{Z}/\langle n \rangle$.*

Example 5.1.16 *Consider the sets* $G = \{e, a, b, c\}$ *and* $G_1 = \{1, -1, i, -i\}$. *Define* $*$ *and* $\cdot$ *on* G *and* G_1, *respectively, by means of the following operation tables.*

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

$\cdot$	1	-1	i	$-i$
1	1	-1	i	$-i$
-1	-1	1	$-i$	i
i	i	$-i$	-1	1
$-i$	$-i$	i	1	-1

Now G_1 *is a cyclic group generated by* i. G *is also a group. However, since* $aa = e$, $bb = e$, *and* $cc = e$, *no element of* G *has order* 4 *and so* G *is not cyclic. Thus,* G *and* G_1 *are not isomorphic.*

Example 5.1.17 *Let* $(\mathbb{R}, +)$ *be the group of real numbers under addition and* $(\mathbb{R}^+, \cdot)$ *be the group of positive real numbers under multiplication. Define* $f : \mathbb{R} \to \mathbb{R}^+$ *by* $f(a) = e^a$ *for all* $a \in \mathbb{R}$. *Clearly* f *is well defined. Let* $a, b \in \mathbb{R}$. *Then* $f(a + b) = e^{a+b} = e^a e^b = f(a)f(b)$. *Hence,* f *is a homomorphism. Suppose* $f(a) = f(b)$. *Then* $e^a = e^b$ *and so* $\log_e e^a = \log_e e^b$. *This implies that* $a = b$, *whence* f *is one-one. Let* $b \in \mathbb{R}^+$. *Then* $\log_e b \in \mathbb{R}$ *and* $f(\log_e b) = e^{\log_e b} = b$. *Thus,* f *is onto* $\mathbb{R}^+$. *Consequently,* f *is an isomorphism of* $(\mathbb{R}, +)$ *onto* $(\mathbb{R}^+, \cdot)$.

Example 5.1.18 *Consider the groups* $(\mathbb{Z}, +)$ *and* $(\mathbb{Q}, +)$. *By Worked-Out Exercise* 1 *(page* 80*),* $(\mathbb{Q}, +)$ *is not cyclic. Since* $(\mathbb{Z}, +)$ *is cyclic and* $(\mathbb{Q}, +)$ *is not cyclic,* $(\mathbb{Z}, +)$ *is not isomorphic to* $(\mathbb{Q}, +)$ *by Theorem* 5.1.14(v).

Example 5.1.19 *The group* $(\mathbb{Q}, +)$ *is not isomorphic to* $(\mathbb{Q}^*, \cdot)$ *since every nonidentity element of* $(\mathbb{Q}, +)$ *is of infinite order while* -1 *is a nonidentity element of* $(\mathbb{Q}^*, \cdot)$ *which is of finite order.*

Let us now characterize finite and infinite cyclic groups.

Theorem 5.1.20 *Every finite cyclic group of order* n *is isomorphic to* $(\mathbb{Z}_n, +_n)$ *and every infinite cyclic group is isomorphic to* $(\mathbb{Z}, +)$.

Proof. Let $(\langle a \rangle, *)$ be a cyclic group of order n. Let $G = \langle a \rangle$. Define the function $f : G \to \mathbb{Z}_n$ by for all $a^i \in G$, $f(a^i) = [i]$. Now $a^i = a^j$ if and only if $a^{j-i} = e$ if and only if $n \mid (j - i)$ if and only if $[i] = [j]$ (Exercise 11, page 23) if and only if $f(a^i) = f(a^j)$. Thus, f is a one-one function. Now

$$f(a^i a^j) = f(a^{i+j}) = [i + j] = [i] +_n [j] = f(a^i) +_n f(a^j).$$

Since f is one-one and G and $\mathbb{Z}_n$ are finite with same number of elements, f is onto $\mathbb{Z}_n$. Hence, $G \simeq \mathbb{Z}_n$.

Now let $G = \langle a \rangle$ be an infinite cyclic group. Define the function$f : G \to \mathbb{Z}$ by $f(a^i) = i$ for all $i \in \mathbb{Z}$. Since $a^i = a^j$ if and only if $a^{i-j} = e$ if and only if $i - j = 0$ (since a is of infinite order) if and only if $i = j$, we have that f is a one-one function of G into $\mathbb{Z}$. From the definition of f, f is onto $\mathbb{Z}$. Now

$$f(a^i a^j) = f(a^{i+j}) = i + j = f(a^i) + f(a^j).$$

Hence, $G \simeq \mathbb{Z}$. ■

Corollary 5.1.21 *Any two cyclic groups of the same order are isomorphic.* ■

From the above corollary, it follows that there is only one (up to isomorphism) cyclic group having a prescribed order.

In Example 5.1.16, we saw that there are at least two nonisomorphic groups of order 4. We now show that these are exactly two nonisomorphic groups of order 4.

Let G be a group of order 4 which is not cyclic. (Example 5.1.16 shows that such a group exists.) Then no element of G can have order 4, for if $a \in G$ has order 4, then e, a, a^2, a^3 would be distinct elements of G and thus G would be cyclic, i.e., $G = \langle a \rangle$. This is contrary to the assumption that G is not cyclic. Let $G = \{e, a, b, c\}$. Since the order of every element of G divides the order of G, a, b, and c have order 2. If $ab = a$, then $b = e$, a contradiction. Thus, $ab \neq a$. Similarly, $ab \neq b$. Suppose $ab = e$, then $a(ab) = ae$. Therefore, $b = a$ since $a^2 = e$, a contradiction. Thus, $ab = c$. Similarly, $ba = c$. Hence, $ab = ba$. By similar arguments, we have $ac = b = ca$

and $bc = a = cb$. Thus, we find that G is a commutative group and its operation table is given by the table in Example 5.1.16. Consequently, there is essentially one group of order 4 which is not cyclic. This is the Klein 4-group. Since all cyclic groups of the same orders are isomorphic, we thus have exactly two nonisomorphic groups of order 4, namely, the Klein 4-group and the cyclic group of order 4. We have thus proved the following result.

Theorem 5.1.22 *There are only two groups of order* 4 *(up to isomorphism), a cyclic group of order* 4 *and* K_4 *(Klein* 4*-group).*

Since every cyclic group is commutative and every group of prime order is cyclic, it follows that that if a group is noncommutative, then it must have order at least 6. Indeed, the symmetric group S_3 is noncommutative and of order 6. Since all cyclic groups of the same order are isomorphic and since every group of prime order is cyclic, there is exactly one group of order 1, 2, 3, 5 (up to isomorphism), respectively. We have seen that there are two nonisomorphic groups of order 4. In the next theorem, we show that there are only two (up to isomorphism) nonisomorphic groups of order 6.

Theorem 5.1.23 *There are only two (up to isomorphism) groups of order* 6.

Proof. The group $\mathbb{Z}_6$ is a cyclic group of order 6 and S_3 is a noncommutative group of order 6. Note that $\mathbb{Z}_6$ is not isomorphic to S_3. To show that there are only two (up to isomorphism) nonisomorphic groups of order 6, we will show that any group of order 6 is isomorphic to either $\mathbb{Z}_6$ or S_3.

Let G be a group of order 6. Since $|G|$ is even, there exists $a \in G$, $a \neq e$ such that $a^2 = e$. If $x^2 = e$ for all $x \in G$, then G is commutative and for any two distinct nonidentity elements a and b, $\{e, a, b, ab\}$ is a subgroup of G. Since $|G| = 6$, G has no subgroups of order 4. Hence, there exists $b \in G$ such that $b^2 \neq e$, i.e., $b \neq e$ and $\circ(b) \neq 2$. Since $\circ(b) \mid 6$, $\circ(b) = 6$ or 3. If $\circ(b) = 6$, then $G = \langle b \rangle$ is a cyclic group of order 6 and $G \simeq \mathbb{Z}_6$. Suppose G is not cyclic. Then $\circ(b) = 3$. Let $H = \{e, b, b^2\}$. Then H is a subgroup of G of index 2. Thus, H is a normal subgroup of G. Clearly $a \notin H$. Now $G = H \cup aH$ and $H \cap aH = \emptyset$. Hence, $G = \{e, b, b^2, a, ab, ab^2\}$. Now $aba^{-1} \in H$ since H is normal and $b \in H$. Therefore, $aba^{-1} = e$ or $aba^{-1} = b$ or $aba^{-1} = b^2$. If $aba^{-1} = e$, then $b = e$, which is a contradiction. If $aba^{-1} = b$, then $ab = ba$. Since $\circ(a)$ and $\circ(b)$ are relatively prime and $ab = ba$, $\circ(ab) = \circ(a) \cdot \circ(b) = 6$. Thus, G is cyclic, a contradiction. Hence, $aba^{-1} = b^2$. Thus, $G = \langle a, b \rangle$, where $\circ(a) = 2$, $\circ(b) = 3$, and $aba^{-1} = b^2$. It is now easy to see that $G \simeq S_3$. ■

We conclude this section by proving Cayley's theorem, which says that any group can be realized as a permutation group.

Theorem 5.1.24 (Cayley) *Any group* G *is isomorphic to some subgroup of the group* $(S(G), \circ)$ *of all permutations of the set* G.

Proof. Let a be an element of a group G. Define the function $f_a : G \to G$ by for all $b \in G$, $f_a(b) = ab$. Then $b = c$ if and only if $ab = ac$ if and only if $f_a(b) = f_a(c)$. Thus, f_a is a one-one function of G into G. For any $b \in G$,

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

So we find that f_a maps G onto G. Hence, f_a is a permutation of G. This implies that $f_a \in S(G)$. Let $F(G) = \{f_a \mid a \in G\}$. Then $F(G)$ is a subset of the set $S(G)$ of all permutations on G Define $g : G \to S(G)$ by for all $a \in G$, $g(a) = f_a$. Then $a = b$ if and only if $ac = bc$ for all $c \in G$ if and only if $f_a(c) = f_b(c)$ for all $c \in G$ if and only if $f_a = f_b$ if and only if $g(a) = g(b)$. This proves that g is a one-one function of G into $F(G)$. Clearly g maps G onto $F(G)$. Now $g(ab) = f_{ab}$ and $g(a) \circ g(b) = f_a \circ f_b$. Also, for all $c \in G$,

$$f_{ab}(c) = (ab)c = a(bc) = f_a(bc) = f_a(f_b(c)) = (f_a \circ f_b)(c).$$

Thus, $f_{ab} = f_a \circ f_b$. Hence, $g(ab) = g(a) \circ g(b)$ and so g is a homomorphism.This implies that $F(G)$ is a subgroup and G is isomorphic to this subgroup. ■

Cayley's theorem is another example of a representation theorem. However, Cayley realized that the best way of studying general problems in group theory was not necessarily by the use of permutations.

Worked-Out Exercises

♢ **Exercise 1** Let $f : G \to G_1$ be an epimorphism of groups. If H is a normal subgroup of G, then show that $f(H)$ is a normal subgroup of G_1.

Solution: By Theorem 5.1.2, we find that $f(H)$ is a subgroup of G_1. Let $g_1 \in G_1$. Since f is onto G_1, there exists $g \in G$ such that $f(g) = g_1$. Let $a \in g_1 f(H) g_1^{-1} = f(g)f(H)f(g)^{-1}$. Then $a = f(g)f(h)f(g)^{-1} = f(ghg^{-1})$ for some $h \in H$. Since H is a normal subgroup of G, $ghg^{-1} \in H$ and so $a \in f(H)$. Thus, $g_1 f(H) g_1^{-1} \subseteq f(H)$. Hence, $f(H)$ is a normal subgroup of G_1.

♢ **Exercise 2** Let G and H be finite groups such that $\gcd(|G|, |H|) = 1$. Show that the trivial homomorphism is the only homomorphism from G into H.

Solution: Let $f : G \to H$ be a homomorphism and let $a \in G$. We show that every element of G is mapped onto the identity element of H, i.e., $f(a) = e_H$ for all $a \in G$, where e_H denotes the identity element of H. Now $\circ(a) \mid |G|$ and $\circ(f(a)) \mid |H|$. Also, by Theorem 5.1.2, $\circ(f(a)) \mid \circ(a)$. Hence, $\circ(f(a)) \mid |G|$. Since $|G|$ and $|H|$ are relatively prime, $\circ(f(a)) = 1$, proving $f(a) = e_H$. Thus, f is the trivial homomorphism.

♢ **Exercise 3** Show that the group $(\mathbb{Q}, +)$ is not isomorphic to $(\mathbb{Q}/\mathbb{Z}, +)$.

Solution: In $(\mathbb{Q}, +)$, every nonzero element is of infinite order. Let $\frac{p}{q} + \mathbb{Z} \in \mathbb{Q}/\mathbb{Z}$, where $p, q \in \mathbb{Z}$ and $q \neq 0$. Then $q(\frac{p}{q} + \mathbb{Z}) = p + \mathbb{Z} = \mathbb{Z}$. This shows that every element of $\mathbb{Q}/\mathbb{Z}$ is of finite order. Hence, $(\mathbb{Q}, +)$ is not isomorphic to $(\mathbb{Q}/\mathbb{Z}, +)$.

Exercise 4 Show that $\mathbb{R}^*$, the group of all nonzero real numbers under multiplication, is not isomorphic to $\mathbb{C}^*$, the group of all nonzero complex numbers under multiplication.

Solution: In the group $\mathbb{C}^*$, i is an element of order 4. But $\mathbb{R}^*$ does not contain any element of order 4. Hence, by Theorem 5.1.14, $\mathbb{R}^*$ is not isomorphic to $\mathbb{C}^*$.

♢ **Exercise 5** Find all homomorphisms from $\mathbb{Z}_6$ into $\mathbb{Z}_4$.

Solution: $\mathbb{Z}_6 = \langle [1] \rangle$. Let $f : \mathbb{Z}_6 \to \mathbb{Z}_4$ be a homomorphism. For any $[a] \in \mathbb{Z}_6$, $f([a]) = af([1])$ shows that f is completely known if $f([1])$ is known. Now $\circ(f([1]))$ divides $\circ([1])$ and 4, i.e., $\circ(f([1]))$ divides 6 and 4. Hence, $\circ(f([1])) = 1$ or 2. Thus, $f([1]) = [0]$ or $[2]$. If $f([1]) = [0]$, then f is the trivial homomorphism which maps every element to $[0]$. On the other hand, $f([1]) = [2]$ implies that $f([a]) = [2a]$ for all $[a] \in \mathbb{Z}_6$. Thus, $f([a] + [b]) = f([a + b]) = [2(a + b)] = [2a + 2b] = [2a] + [2b] = f([a]) + f([b])$, proving that the mapping $f : \mathbb{Z}_6 \to \mathbb{Z}_4$ defined by $f([a]) = [2a]$ for all $[a] \in \mathbb{Z}_6$ is a homomorphism. Hence, there are two homomorphisms from $\mathbb{Z}_6$ into $\mathbb{Z}_4$.

Exercise 6 Let G be a finite commutative group. Let $n \in \mathbb{Z}$ be such that n and $|G|$ are relatively prime. Show that the function $\phi : G \to G$ defined by $\phi(a) = a^n$ for all $a \in G$ is an isomorphism of G onto G.

Solution: Let $a, b \in G$. Now

$$\begin{aligned} \phi(ab) &= (ab)^n \\ &= a^n b^n \qquad \text{(since } G \text{ is commutative)} \\ &= \phi(a)\phi(b). \end{aligned}$$

This implies that ϕ is a homomorphism. Let $\phi(a) = \phi(b)$. Then $a^n = b^n$ and so $(ab^{-1})^n = e$. Therefore, $\circ(ab^{-1})$ divides n. Since $\circ(ab^{-1})$ divides $|G|$ and n and $|G|$ are relatively prime, $\circ(ab^{-1}) = 1$. This implies that $ab^{-1} = e$, i.e., $a = b$, proving that ϕ is one-one. Since G is a finite group and ϕ is one-one, ϕ is onto G. Hence, ϕ is an isomorphism of G onto G.

♢ **Exercise 7**

(a) Let G be a group and $f : G \to G$ be defined by $f(a) = a^n$ for all $a \in G$, where n is a positive integer. Suppose f is an isomorphism. Prove that $a^{n-1} \in Z(G)$ for all $a \in G$.

(b) Let G be a group and $f : G \to G$ defined by for all $a \in G$, $f(a) = a^3$ be an isomorphism. Prove that G is commutative.

Solution: (a) Let $a, b \in G$. Then $f(a^{-1}ba) = (a^{-1}ba)^n = a^{-1}b^n a$. Thus,

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

Hence, $a^{-(n-1)}\ b^n\ a^{n-1} = b^n$ or $(a^{-(n-1)}ba^{n-1})^n = b^n$. Thus, $f(a^{-(n-1)}ba^{n-1}) = f(b)$. Since f is one-one, $a^{-(n-1)}ba^{n-1} = b$. Hence, $a^{n-1}b = ba^{n-1}$, proving that $a^{n-1} \in Z(G)$.

(b) By (a), $a^2 \in Z(G)$ for all $a \in G$. Let $a, b \in G$. Then $f(ab) = (ab)^3 = ab(ab)^2 = a(ab)^2 b = aababb = a^2bab^2 = ba^2b^2a = bb^2a^2a = b^3a^3 = f(b)f(a) = f(ba)$. Hence, $ab = ba$ since f is one-one. Thus, G is commutative.

Exercises

1. Determine whether the indicated function f is a homomorphism from the first group into the second group. If f is a homomorphism, determine its kernel.

(a) $f(a) = a^2$; $(\mathbb{R}^+, \cdot)$, $(\mathbb{R}^+, \cdot)$ for all $a \in \mathbb{R}^+$.

(b) $f(a) = 2^a$; $(\mathbb{R}, +)$, $(\mathbb{R}^+, \cdot)$ for all $a \in \mathbb{R}$.

(c) $f(a) = |a|$; $(\mathbb{R}\backslash\{0\}, \cdot)$, $(\mathbb{R}^+, \cdot)$ for all $a \in \mathbb{R}\backslash\{0\}$.

(d) $f(a) = a + 1$; $(\mathbb{Z}, +)$, $(\mathbb{Z}, +)$ for all $a \in \mathbb{Z}$.

(e) $f(a) = 2a$; $(\mathbb{Z}, +)$, $(\mathbb{Z}, +)$ for all $a \in \mathbb{Z}$.

(f) $f([a]) = [5a], (\mathbb{Z}_8, +)$, $[\mathbb{Z}_8, +]$

2. Find all homomorphisms from $\mathbb{Z}$ into $\mathbb{Z}$. How many homomorphisms are onto?

3. Find all homomorphisms from $\mathbb{Z}$ onto $\mathbb{Z}_6$.

4. Find all homomorphisms from $\mathbb{Z}_8$ into $\mathbb{Z}_{12}$ and from $\mathbb{Z}_{20}$ into $\mathbb{Z}_{10}$.

5. Show that $\mathbb{Q}^*$, the group of all nonzero rational numbers under multiplication, is not isomorphic to $\mathbb{R}^*$, the group of all nonzero real numbers under multiplication.

6. Show that $(\mathbb{Q}, +)$ is not isomorphic to $(\mathbb{R}, +)$.

7. Show that $(\mathbb{Z}, +)$ is not isomorphic to $(\mathbb{R}, +)$.

8. Let G be a group. Define the function $f : G \to G$ by for all $a \in G$, $f(a) = a^{-1}$. Prove that f is a homomorphism if and only if G is commutative.

9. Let $G = \{(a, b) \mid a, b \in \mathbb{R},\ b \neq 0\}$. Then $(G, *)$ is a noncommutative group under the binary operation $(a, b)*(c, d) = (a+bc, bd)$ for all $(a, b), (c, d) \in G$. Let $H = \{(a, b) \in G \mid a = 0\}$ and $K = \{(a, b) \in G \mid b > 0\}$. Show that $H \cap K \simeq (\mathbb{R}^+, \cdot)$, where $(\mathbb{R}^+, \cdot)$ is the group of all positive real numbers under multiplication.

10. Let $G = \{a \in \mathbb{R} \mid -1 < a < 1\}$. Show that $(G, *) \simeq (\mathbb{R},+)$, where the binary operation $*$ on G is defined by

$$a * b = \frac{a + b}{1 + ab}$$

for all $a, b \in G$.

11. (a) Let f be a homomorphism from a cyclic group of order 8 onto a cyclic group of order 4. Determine Ker f.

(b) Let f be a homomorphism from a cyclic group of order 8 onto a cyclic group of order 2. Determine Ker f.

12. Prove that a homomorphic image of a cyclic group is cyclic.

13. Show that S_3 and $\mathbb{Z}_6$ are not isomorphic groups, but for every proper subgroup A of S_3 there exists a proper subgroup B of $\mathbb{Z}_6$ such that $A \simeq B$.

14. Let G, H, and K be groups. Suppose that the functions $f : G \to H$ and $g : H \to K$ are homomorphisms. Prove that $g \circ f : G \to K$ is also a homomorphism.

15. Let G and H be groups. Define the function $f : G \times H \to G$ by for all $(a, b) \in G \times H$, $f((a, b)) = a$. Prove that f is a homomorphism from $G \times H$ onto G. Determine Ker f.

16. Let $f : G \to H$ be an isomorphism of groups. Prove that $f^{-1} : H \to G$ is also an isomorphism.

17. Let G, H, and K be groups. Prove that

(a) $G \times H \simeq H \times G$.

(b) If $G \simeq H$ and $H \simeq K$, then $G \simeq K$.

(c) $G \times (H \times K) \simeq (G \times H) \times K$.

18. Let G and H be groups. Let $f : G \to H$ be a homomorphism of G onto H. Show that if $G = \langle S \rangle$ for some subset S of G, then $H = \langle f(S) \rangle$.

19. Let $f : G \to H$ be an isomorphism of groups. Show that for any integer k and for any $g \in G$, the sets $A = \{a \in G \mid a^k = g\}$ and $B = \{b \in H \mid b^k = f(g)\}$ have the same number of elements.

20. Let G be a simple group and $\psi : S_n \to G$ be an epimorphism for some positive integer n. Prove that $G \simeq S_k$ for some $k \leq n$.

21. Which of the following statements are true? Justify.

(a) A cyclic group with more than one element may be a homomorphic image of a noncyclic group.

(b) There does not exist a nontrivial homomorphism from a group G of order 5 into a group H of order 4.

(c) The group $(\mathbb{Z},+)$ is isomorphic to $(\mathbb{Q},+)$.

(d) There exists a monomorphism from a group of order 20 into a group of order 70.

(e) There exists an epimorphism of $(\mathbb{R},+)$ onto $(\mathbb{Z},+)$.

(f) There does not exist any epimorphism of $(\mathbb{Q},+)$ onto $(\mathbb{Z},+)$.

(g) If f and g are two epimorphisms of a group G onto a group H such that Ker f = Ker g, then $f = g$.

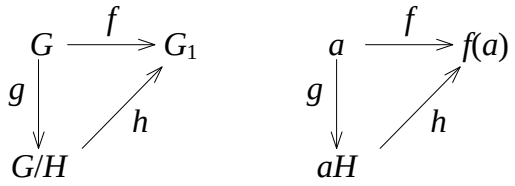
(h) $(\mathbb{Z} \times \mathbb{Z}, +)$ is a cyclic group.

(i) The group $(\mathbb{Z}, +)$ is a homomorphic image of $(\mathbb{Q}, +)$.

5.2 Isomorphism and Correspondence Theorems

In this section, we continue our study of isomorphisms. Our objective is to prove the fundamental theorem of homomorphisms, the isomorphism theorems, and the correspondence theorem. These theorems show us the relationship between homomorphisms and quotient groups.

Theorem 5.2.1 *Let f be a homomorphism of a group G onto a group G_1, H be a normal subgroup of G such that $H \subseteq$ Ker f, and g be the natural homomorphism of G onto G/H. Then there exists a unique homomorphism h of G/H onto G_1 such that $f = h \circ g$. Furthermore, h is one-one if and only if $H =$ Ker f.*



Proof. Define $h : G/H \to G_1$ by

$$h(aH) = f(a)$$

for all $aH \in G/H$.

Now $aH = bH$ implies $b^{-1}a \in H \subseteq$ Ker f and so $f(b^{-1}a) = e_1$ or $f(a) = f(b)$. Hence, $h(aH) = h(bH)$ and so h is well defined. Let $a \in G$. Then

$$(h \circ g)(a) = h(g(a)) = h(aH) = f(a).$$

Therefore, $h \circ g = f$. Since f maps G onto G_1, h must map G/H onto G_1. Now

$$h((aH)(bH)) = h((ab)H) = f(ab) = f(a)f(b) = h(aH)h(bH).$$

Hence, h is a homomorphism of G/H onto G_1 satisfying $f = h \circ g$. To prove the uniqueness part, let us assume $f = h' \circ g$ for some homomorphism h' from G/H onto G_1. Then

$$h(aH) = f(a) = (h' \circ g)(a) = h'(g(a)) = h'(aH)$$

for all $aH \in G/H$ and so $h = h'$. Hence, h is the only homomorphism of G/H onto G_1 such that $f = h \circ g$.

Suppose h is one-one. Let $a \in$ Ker f. Then $f(a) = e_1$ and so $h(aH) = e_1$. Since $h(eH) = e_1$ and h is one-one, $aH = eH$. Thus, $a \in H$ and so Ker $f \subseteq H$. By hypothesis, $H \subseteq$ Ker f and so $H =$ Ker f. Conversely, assume $H =$ Ker f. Suppose $h(aH) = h(bH)$. Then $f(a) = f(b)$ or $f(b^{-1}a) = e_1$. Thus, $b^{-1}a \in$ Ker $f = H$ and so $aH = bH$, proving that h is one-one. ■

From Theorem 5.2.1, it follows that if $H =$ Ker f, then h is an isomorphism and hence $G/$Ker f is isomorphic to G_1, i.e., every homomorphism of a group G onto a group G_1 induces an isomorphism of $G/$Ker f onto G_1. This result plays a fundamental role in group theory. It is known as **the fundamental theorem of homomorphisms** for groups. This result is also called the first isomorphism theorem for groups. Considering the importance of this theorem, we state it in its general form and also give a direct proof of it.

Theorem 5.2.2 (First Isomorphism Theorem) *Let f be a homomorphism of a group G into a group G_1. Then $f(G)$ is a subgroup of G_1 and*

$$G/Ker\ f \simeq f(G).$$

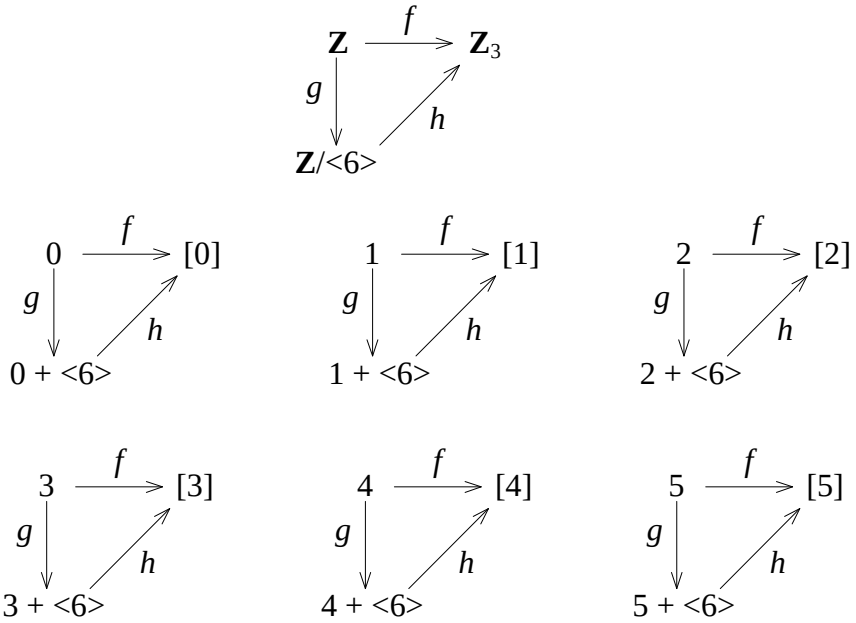
Proof. By Theorem 5.1.2, $f(G)$ is a subgroup of G_1. Let $H = \text{Ker } f$. Define $h : G/H \to f(G)$ by

$$h(aH) = f(a)$$

for all $aH \in G/H$. Now $aH = bH$ if and only if $b^{-1}a \in H = \text{Ker } f$ if and only if $f(b^{-1}a) = e_1$ if and only if $f(b^{-1})f(a) = e_1$ if and only if $f(a) = f(b)$. Thus, h is a one-one function. Let $x \in f(G)$. Then $x = f(b)$ for some $b \in G$. Therefore, $h(bH) = f(b) = x$. This shows that h is onto $f(G)$. Finally, $h(aHbH) = h(abH) = f(ab) = f(a)f(b) = h(aH)h(bH)$ for all $aH, bH \in G/H$, proving that h is a homomorphism. Consequently, $G/\text{Ker } f \simeq f(G)$. ■

In the following example we illustrate the first isomorphism theorem.

Example 5.2.3 *Let f be the homomorphism of $(\mathbb{Z}, +)$ onto $(\mathbb{Z}_3, +_3)$ defined by $f(n) = [n]$ for all $n \in \mathbb{Z}$. Let g be the natural homomorphism of $\mathbb{Z}$ onto $\mathbb{Z}/\langle 6\rangle$. Now $\langle 6\rangle$ is a normal subgroup of $\mathbb{Z}$ and $\langle 6\rangle \subset \langle 3\rangle = Ker\ f$. Thus, there exists a homomorphism h of $\mathbb{Z}/\langle 6\rangle$ onto $\mathbb{Z}_3$ such that $f = h \circ g$. The homomorphism h is defined by $h(n + \langle 6\rangle) = [n]$.*



Recall that a group G_1 is called a **homomorphic image** of a group G if there exists a homomorphism of G onto G_1.

From Theorem 5.2.1 and Corollary 5.2.2, we find that for each normal subgroup N of a group G, G/N is a homomorphic image of G, and for each homomorphic image G_1, there exists a normal subgroup N of G such that $G/N \simeq G_1$.

Example 5.2.4 *The group S_3 has (up to isomorphism) only three homomorphic images. This follows from the fact that S_3 has only three normal subgroups. The homomorphic images are S_3, $\mathbb{Z}_1$, and $\mathbb{Z}_2$ since $\{e\}$, S_3, and $\{e, (1\ 2\ 3), (1\ 3\ 2)\}$ are the only normal subgroups of S_3 and $S_3 \simeq S_3/\{e\}$, $\mathbb{Z}_1 \simeq S_3/S_3$, and $\mathbb{Z}_2 \simeq S_3/\{e, (1\ 2\ 3), (1\ 3\ 2)\}$.*

Theorem 5.2.5 *Let G_1 be a homomorphic image of a group G. Then the following assertions hold.*

(i) If G is cyclic, then G_1 is cyclic.

(ii) If G is commutative, then G_1 is commutative.

(iii) If G_1 contains an element of order n and $|G|$ is finite, then G contains an element of order n.

Proof. (i) Follows by Exercise 12 (page 104).

(ii) Follows by Theorem 5.1.2(v).

(iii) Let $f : G \to G_1$ be an epimorphism and let a' be an element of G_1 of order n. If $n = 1$, then e is the required element of G of order 1. Suppose $n > 1$. Since f is onto G_1, there exists $a \in G$ such that $f(a) = a'$. Now $\circ(a)$ is finite and by Theorem 5.1.2(v), $\circ(a')$ divides $\circ(a)$, i.e., n divides $\circ(a)$. Let $t \in \mathbb{Z}^+$ be such that $\circ(a) = nt$. Then $t < \circ(a)$. Hence, $a^t \neq e$. Now $a^{nt} = e$. Let $b = a^t$. Then $b^n = e$ and by Theorem 2.1.46,

$$\circ(a^t) = \frac{\circ(a)}{\gcd(t, \circ(a))} = \frac{nt}{t} = n.$$

■

Note that the result in Theorem 5.2.5(iii) does not hold if $|G|$ is not finite. For example, $\mathbb{Z}_6$ is a homomorphic image of $\mathbb{Z}$; $\mathbb{Z}_6$ contains an element of order 3, but $\mathbb{Z}$ has no element of order 3.

Theorem 5.2.6 (Second Isomorphism Theorem) *Let H and K be subgroups of a group G with K normal in G. Then*

$$H/(H \cap K) \simeq (HK)/K.$$

Proof. Define $f : H \to (HK)/K$ by $f(h) = hK$ for all $h \in H$. Now

$$f(h_1h_2) = h_1h_2K = h_1Kh_2K = f(h_1)f(h_2)$$

for all $h_1, h_2 \in H$, proving that f is a homomorphism. Let $xK \in (HK)/K$. Then $x = hk$ for some $h \in H$ and $k \in K$. Thus,

$$xK = (hk)K = (hK)(kK) = hK = f(h).$$

This proves that f is onto $(HK)/K$ and so $f(H) = (HK)/K$. Hence, by the first isomorphism theorem, it follows that

$$H/\text{Ker } f \simeq (HK)/K.$$

To complete the proof, we show that Ker $f = H \cap K$. Now

$$\begin{aligned} \text{Ker } f &= \{h \in H \mid f(h) = \text{ identity element of } HK/K\} \\ &= \{h \in H \mid hK = K\} \\ &= \{h \in H \mid h \in K\} \\ &= H \cap K. \end{aligned}$$

Consequently, $H/H \cap K \simeq (HK)/K$. ■

We illustrate the second isomorphism theorem with the help of the following example.

Example 5.2.7 *Consider the group $(\mathbb{Z}, +)$ and its subgroups $H = \langle 2\rangle$ and $K = \langle 3\rangle$. Then $H+K = \langle 2\rangle + \langle 3\rangle = \mathbb{Z}$ and $H \cap K = \langle 6\rangle$. Theorem 5.2.6 says that*

$$H/(H \cap K) \simeq (H + K)/K,$$

i.e.,

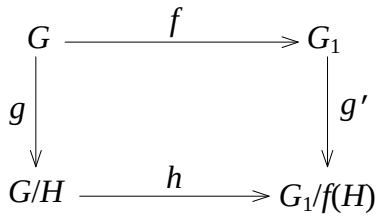
$$\langle 2\rangle / \langle 6\rangle \simeq \mathbb{Z}/\langle 3\rangle.$$

This isomorphism is evident if we notice that $\langle 2\rangle / \langle 6\rangle = \{0 + \langle 6\rangle, 2 + \langle 6\rangle, 4 + \langle 6\rangle\}$ while $\mathbb{Z}/\langle 3\rangle = \{0 + \langle 3\rangle, 1 + \langle 3\rangle, 2 + \langle 3\rangle\}$. The mapping

$$h : \langle 2\rangle / \langle 6\rangle \to \mathbb{Z}/\langle 3\rangle$$

defined by $h : 0 + \langle 6\rangle \to 0 + \langle 3\rangle$, $2 + \langle 6\rangle \to 2 + \langle 6\rangle \to\!\!\to \langle 3\rangle$, $4 + \langle 6\rangle \to 1 + \langle 3\rangle$ is the desired isomorphism.

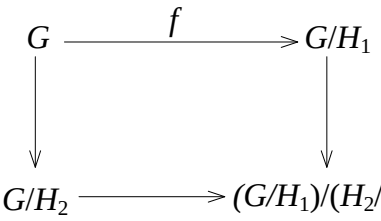
Theorem 5.2.8 *Let f be a homomorphism of a group G onto a group G_1, H be a normal subgroup of G such that $H \supseteq$ Ker f, and g, g' be the natural homomorphisms of G onto G/H and G_1 onto $G_1/f(H)$, respectively. Then there exists a unique isomorphism h of G/H onto $G_1/f(H)$ such that $g' \circ f = h \circ g$.*



Proof. If we show Ker $g' \circ f = H$, then there exists a unique isomorphism h of G/H onto $G_1/f(H)$ by Theorem 5.2.1. Let $a \in H$. Then $(g' \circ f)(a) = g'(f(a)) =$ the identity of $G_1/f(H)$ since $f(a) \in f(H) =$ Ker g'. Thus, $a \in$ Ker $g' \circ f$ and hence $H \subseteq$ Ker $g' \circ f$. Let $a \in$ Ker $g' \circ f$. Then $g'(f(a)) =$ the identity of $G_1/f(H)$ and so $f(a) \in$ Ker $g' = f(H)$. Therefore, there exists $b \in H$ such that $f(b) = f(a)$ or $f(ab^{-1}) = e_1$. This implies that $ab^{-1} \in$ Ker $f \subseteq H$ and so $a = (ab^{-1})b \in H$. Thus, Ker $g' \circ f \subseteq H$. Hence, Ker $g' \circ f = H$. ■

Corollary 5.2.9 (Third Isomorphism Theorem) *Let H_1, H_2 be normal subgroups of a group G such that $H_1 \subseteq H_2$. Then*

$$(G/H_1)/(H_2/H_1) \simeq G/H_2.$$



Proof. Make the following substitutions in Theorem 5.2.8: G/H_1 for G_1, H_2 for H, and $(G/H_1)/(H_2/H_1)$ for $G_1/f(H)$, where in this case f is the natural homomorphism of G onto G/H_1. Note that $f(H_2) = H_2/H_1$. ■

We illustrate the third isomorphism theorem with the help of the following example.

Example 5.2.10 *Consider the group $(\mathbb{Z}, +)$ and the subgroups $\langle 6\rangle$ and $\langle 3\rangle$ of $\mathbb{Z}$. Then*

$$\mathbb{Z}/\langle 3\rangle = \{0 + \langle 3\rangle, 1 + \langle 3\rangle, 2 + \langle 3\rangle\}.$$

$$\begin{aligned} \mathbb{Z}/\langle 6\rangle &= \{0 + \langle 6\rangle, 1 + \langle 6\rangle, 2 + \langle 6\rangle, 3 + \langle 6\rangle, 4 + \langle 6\rangle, 5 + \langle 6\rangle\}. \\ \langle 3\rangle/\langle 6\rangle &= \{0 + \langle 6\rangle, 3 + \langle 6\rangle\}. \end{aligned}$$

Now,

$$(\mathbb{Z}/\langle 6\rangle)/(\langle 3\rangle/\langle 6\rangle) = \{\bar{0}, \bar{1}, \bar{2}\},$$

where

$$\begin{aligned} \bar{0} &= 0 + \langle 6\rangle + (\langle 3\rangle/\langle 6\rangle) \\ \bar{1} &= 1 + \langle 6\rangle + (\langle 3\rangle/\langle 6\rangle) \\ \bar{2} &= 2 + \langle 6\rangle + (\langle 3\rangle/\langle 6\rangle). \end{aligned}$$

It is now clear that

$$\mathbb{Z}/\langle 3\rangle \simeq (\mathbb{Z}/\langle 6\rangle)/(\langle 3\rangle/\langle 6\rangle)$$

since both are cyclic groups of order 3 *and of course, by Corollary* 5.2.9.

We can at times determine the subgroups of a group G_1 from a group G whose subgroups are known if there is a homomorphism f of G onto G_1. For if such an f exists, the following result says that the subgroups of G_1 can be determined from the subgroups of G which contain Ker f.

Theorem 5.2.11 (Correspondence Theorem) *Let f be a homomorphism of a group G onto a group G_1. Then f induces a one-one inclusion preserving correspondence between the subgroups of G containing Ker f and the subgroups of G_1. In fact, if H and K are corresponding subgroups of G and G_1, respectively, then H is a normal subgroup of G if and only if K is a normal subgroup of G_1.*

Proof. Let

$$\mathcal{H} = \{H \mid H \text{ is a subgroup of } G \text{ such that Ker } f \subseteq H\}$$

and

$$\mathcal{K} = \{K \mid K \text{ is a subgroup of } G_1\}.$$

Define $f^* : \mathcal{H} \to \mathcal{K}$ by for all $H \in \mathcal{H}$, $f^*(H) = \{f(h) \mid h \in H\}$. Then $f^*(H) \in \mathcal{K}$ by Theorem 5.1.2. Hence, f^* is a function since f is a function. Let $K \in \mathcal{K}$. Denote the preimage, $f^{-1}(K)$, of K in G by H. Let $a \in$ Ker f. Then $f(a) = e_1 \in K$ and so $a \in f^{-1}(K) = H$. Thus, Ker $f \subseteq H$. Let $a, b \in H$. Then $f(a), f(b) \in K$ and so $f(ab^{-1}) = f(a)f(b^{-1}) = f(a)f(b)^{-1} \in K$. Therefore, $ab^{-1} \in H$ and so H is a subgroup of G containing Ker f, i.e., $H \in \mathcal{H}$. Hence, f^* maps $\mathcal{H}$ onto $\mathcal{K}$. Let $H_1, H_2 \in \mathcal{H}$. Suppose $f^*(H_1) = f^*(H_2)$. Let $h_1 \in H_1$. Then there

exists $h_2 \in H_2$ such that $f(h_1) = f(h_2)$. This implies that $f(h_1 h_2^{-1}) = e_1$ and so $h_1 h_2^{-1} \in \text{Ker } f \subseteq H_2$. Hence, $h_1 = (h_1 h_2^{-1}) h_2 \in H_2$. Therefore, $H_1 \subseteq H_2$. Similarly, $H_2 \subseteq H_1$. Thus, $H_1 = H_2$ and so f^* is one-one. Clearly $H_1 \subseteq H_2$ if and only if $f^*(H_1) \subseteq f^*(H_2)$. In fact, since f^* is one-one, $H_1 \subset H_2$ if and only if $f^*(H_1) \subset f^*(H_2)$.

Suppose H is a normal subgroup of G such that Ker $f \subseteq H$. Let $K = f^*(H)$. We show that K is a normal subgroup of G. Let $f(a) \in G_1$ and $f(h) \in K$. Now $aha^{-1} \in H$ since H is a normal subgroup of G and so $f(a)f(h)f(a)^{-1} = f(aha^{-1}) \in K$. Hence, K is a normal subgroup of G_1. Let J be a normal subgroup of G_1 and $L \in \mathcal{H}$ be such that $f^*(L) = J$. Let $a \in G$ and $h \in L$. Then $f(aha^{-1}) = f(a)f(h)f(a)^{-1} \in J$ and so $aha^{-1} \in L$. This proves that L is a normal subgroup of G. ■

Corollary 5.2.12 *Let N be a normal subgroup of a group G. Then every subgroup of G/N is of the form K/N, where K is a subgroup of G that contains N. Also, K/N is a normal subgroup of G/N if and only if K is a normal subgroup of G.*

Proof. Let $g : G \to G/N$ be the natural homomorphism. If $a \in G$, then $g(a) = aN$. From Theorem 5.2.11, we find that this homomorphism induces a one-one mapping g^* between the subgroups of G which contain Ker $g = N$ and the subgroups of G/N. Let H be a subgroup of G/N. Then there exists a subgroup K of G such that $N \subseteq K$ and $H = g^*(K) = \{g(a) \mid a \in K\} = K/N$. The last part follows from Theorem 5.2.11. ■

The following example illustrates the correspondence theorem.

Example 5.2.13 *Let f be a homomorphism of $(\mathbb{Z}, +)$ onto $(\mathbb{Z}_{12}, +_{12})$ defined by $f(n) = [n]$ for all $n \in \mathbb{Z}$. Then for $\mathcal{H}$ and $\mathcal{K}$ of Theorem* 5.2.11*,*

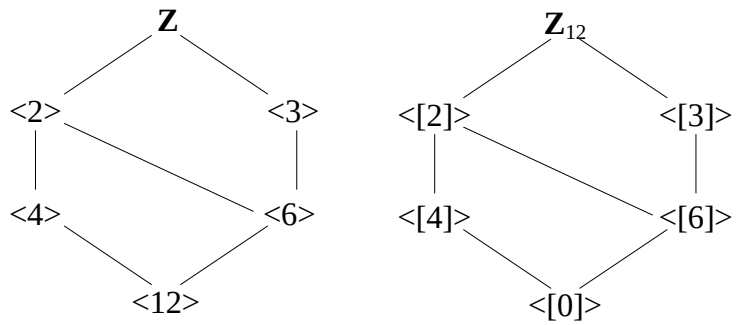
$$\mathcal{H} = \{\langle 12\rangle, \langle 6\rangle, \langle 4\rangle, \langle 3\rangle, \langle 2\rangle, \mathbb{Z}\}$$

and

$$\mathcal{K} = \{\langle [0]\rangle, \langle [6]\rangle, \langle [4]\rangle, \langle [3]\rangle, \langle [2]\rangle, \mathbb{Z}_{12}\}.$$

$$\begin{array}{ll} f^* : \langle 12\rangle \to \langle [0]\rangle, & f^* : \langle 3\rangle \to \langle [3]\rangle, \\ f^* : \langle 2\rangle \to \langle [2]\rangle, & f^* : \langle 6\rangle \to \langle [6]\rangle, \\ f^* : \langle 4\rangle \to \langle [4]\rangle, & f^* : \mathbb{Z} \to \mathbb{Z}_{12}. \end{array}$$

The following diagram indicates the one-one inclusion preserving the correspondence property of f^*.



Now $\langle [9]\rangle = \{n[9] \mid n \in \mathbb{Z}\} \subseteq \{n[3] \mid n \in \mathbb{Z}\} = \langle [3]\rangle$. Also, $[3] = [27] = 3[9] \in \langle [9]\rangle$. Therefore, $\langle [3]\rangle \subseteq \langle [9]\rangle$. Hence, $\langle [3]\rangle = \langle [9]\rangle$. Thus, the subgroup $\langle 9\rangle$ of $\mathbb{Z}$ gets mapped to the subgroup $\langle [3]\rangle$ of $\mathbb{Z}_{12}$ by f. However, this does not contradict Theorem 5.2.11 *since $\langle 9\rangle \not\supseteq \langle 12\rangle$.*

In the remainder of this section, we consider all isomorphisms of a group G onto itself. Recall that Aut(G) is the set of all automorphisms of G.

Theorem 5.2.14 *Let G be a group. Then $(Aut(G), \circ)$ is a group, where $\circ$ denotes the composition of functions.*

Proof. Since $i_G \in \text{Aut}(G)$, Aut$(G) \neq \emptyset$. Let $f, g \in \text{Aut}(G)$. Then $f \circ g$ is an automorphism by Exercise 14 (page 104) and Theorem 1.4.11. Hence, $f \circ g \in \text{Aut}(G)$. Clearly i_G is the identity of Aut(G) and f^{-1} is the inverse of f. Also, $\circ$ is associative by Theorem 1.4.13. Consequently, (Aut$(G), \circ$) is a group. ■

Theorem 5.2.15 *Let G be a group and $a \in G$. Define $\theta_a : G \to G$ by $\theta_a(b) = aba^{-1}$ for all $b \in G$. Then*
(i) $\theta_a \in Aut(G)$,
(ii) $\theta_a \circ \theta_b = \theta_{ab}$ for all $a, b \in G$,
(iii) $(\theta_a)^{-1} = \theta_{a^{-1}}$,
(iv) for all $\alpha \in Aut(G)$, $\alpha \circ \theta_a \circ \alpha^{-1} = \theta_{\alpha(a)}$.

Proof. (i) Let $c, d \in G$. Suppose $c = d$. Then $aca^{-1} = ada^{-1}$ or $\theta_a(c) = \theta_a(d)$. Therefore, θ_a is well defined. Now $\theta_a(cd) = a(cd)a^{-1} = (aca^{-1})(ada^{-1}) = \theta_a(c)\theta_a(d)$. This shows that θ_a is a homomorphism. Also, $c = \theta_a(a^{-1}ca)$, proving that θ_a is onto G. Suppose $\theta_a(c) = \theta_a(d)$. Then $aca^{-1} = ada^{-1}$ and so $c = d$. Thus, θ_a is one-one. Consequently, $\theta_a \in \text{Aut}(G)$.

(ii) Let $a, b \in G$. Then $(\theta_a \circ \theta_b)(c) = \theta_a(\theta_b(c)) = \theta_a(bcb^{-1}) = a(bcb^{-1})a^{-1} = (ab)c(ab)^{-1} = \theta_{ab}(c)$ for all $c \in G$. Hence, $\theta_a \circ \theta_b = \theta_{ab}$.

(iii) Note that $\theta_a \circ \theta_{a^{-1}} = \theta_{aa^{-1}} = \theta_e = i_G$ and $\theta_{a^{-1}} \circ \theta_a = \theta_{a^{-1}a} = \theta_e = i_G$. Thus, $(\theta_a)^{-1} = \theta_{a^{-1}}$.

(iv) Let $\alpha \in \text{Aut}(G)$. Now $(\alpha \circ \theta_a \circ \alpha^{-1})(b) = \alpha(\theta_a(\alpha^{-1}(b))) = \alpha(a\alpha^{-1}(b)a^{-1}) = \alpha(a)\alpha(\alpha^{-1}(b))\alpha(a^{-1}) = \alpha(a)b(\alpha(a))^{-1} = \theta_{\alpha(a)}(b)$ for all $b \in G$. Hence, $\alpha \circ \theta_a \circ \alpha^{-1} = \theta_{\alpha(a)}$. ■

The automorphism θ_a of Theorem 5.2.15 is called an **inner automorphism** of G. We denote by $\text{Inn}(G)$ the set of all inner automorphisms of G.

Theorem 5.2.16 *Let G be a group. Then $Inn(G)$ is a normal subgroup of $Aut(G)$.*

Proof. Since $i_G = \theta_e \in \text{Inn}(G)$, $\text{Inn}(G) \neq \emptyset$. By Theorem 5.2.15(i), $\text{Inn}(G) \subseteq \text{Aut}(G)$. Let $\theta_a, \theta_b \in \text{Inn}(G)$. Then $\theta_a \circ \theta_b^{-1} = \theta_a \circ \theta_{b^{-1}} = \theta_{ab^{-1}} \in \text{Inn}(G)$. Hence, $\text{Inn}(G)$ is a subgroup of $\text{Aut}(G)$ by Theorem 4.1.6. Let $\alpha \in \text{Aut}(G)$. Then by Theorem 5.2.15(iv), $\alpha \circ \theta_a \circ \alpha^{-1} = \theta_{\alpha(a)} \in \text{Inn}(G)$. Hence, $\text{Inn}(G)$ is a normal subgroup of $\text{Aut}(G)$. ■

Theorem 5.2.17 *Let G be a group and H be a subgroup of G. Then*

$$\frac{N(H)}{C(H)} \simeq \textit{ a subgroup of } Aut(H),$$

where $N(H) = \{x \in G \mid xHx^{-1} = H\}$ is the normalizer of H and $C(H) = \{x \in G \mid xhx^{-1} = h \text{ for all } h \in H\}$ is the centralizer of H.

Proof. Define $f : N(H) \to \text{Aut}(H)$ by for all $a \in N(H)$,

$$f(a) = \theta_a \mid_H .$$

Then f is well defined. Let $a_1, a_2 \in N(H)$. Then $f(a_1a_2) = \theta_{a_1a_2}|_H = \theta_{a_1}|_H \circ \theta_{a_2}|_H = f(a_1) \circ f(a_2)$. Thus, f is a homomorphism. Now

$$\begin{aligned}
\text{Ker } f &= \{a \in G \mid f(a) = i_H\} \\
&= \{a \in G \mid \theta_a = i_H\} \\
&= \{a \in G \mid \theta_a(b) = i_H(b) \text{ for all } b \in H\} \\
&= \{a \in G \mid aba^{-1} = b \text{ for all } b \in H\} \\
&= \{a \in G \mid ab = ba \text{ for all } b \in H\} \\
&= C(H).
\end{aligned}$$

Thus, by the first isomorphism theorem, we have the desired result. ■

Corollary 5.2.18 *Let G be a group. Then*

$$\frac{G}{Z(G)} \simeq Inn(G).$$

Proof. Let $H = G$ in Theorem 5.2.17. Then we have $N(G) = G$ and $C(G) = Z(G)$. ■

Worked-Out Exercises

♢ **Exercise 1** Find all homomorphic images of the additive group $\mathbb{Z}$.

Solution: Let H be a homomorphic image of $(\mathbb{Z}, +)$. There exists a homomorphism f of $\mathbb{Z}$ onto H. By the first isomorphism theorem, $\mathbb{Z}/\text{Ker } f \simeq H$. Since Ker f is a subgroup of $\mathbb{Z}$, Ker $f = n\mathbb{Z}$ for some integer $n \geq 0$. Hence, $H \simeq \mathbb{Z}/n\mathbb{Z}$ for some integer $n \geq 0$. On the other hand, for any $n \geq 0$, $n\mathbb{Z}$ is a subgroup of $\mathbb{Z}$ and since $\mathbb{Z}$ is commutative, $n\mathbb{Z}$ is a normal subgroup of $\mathbb{Z}$. There exists a natural homomorphism f from $\mathbb{Z}$ onto $\mathbb{Z}/n\mathbb{Z}$ given by $f(m) = m + n\mathbb{Z}$ for all $m \in \mathbb{Z}$. This shows that $\mathbb{Z}/n\mathbb{Z}$ is a homomorphic image of $\mathbb{Z}$ for all $n \geq 0$. Consequently, the homomorphic images of $\mathbb{Z}$ are the groups (up to isomorphism) $\mathbb{Z}/n\mathbb{Z}$, $n \geq 0$. Now for $n = 0$, $\mathbb{Z}/n\mathbb{Z} \simeq \mathbb{Z}$ and for $n > 0$, $\mathbb{Z}/n\mathbb{Z} \simeq \mathbb{Z}_n$ (Exercise 2, page 111). Therefore, we conclude that the homomorphic images of $\mathbb{Z}$ are the cyclic groups $\mathbb{Z}$ and $\mathbb{Z}_n$, $n > 0$.

♢ **Exercise 2** If there exists an epimorphism of a finite group G onto the group $\mathbb{Z}_8$, show that G has normal subgroups of index 4 and 2.

Solution: Let $f : G \to \mathbb{Z}_8$ be an epimorphism. Then by the first isomorphism theorem, $G/\text{Ker } f \simeq \mathbb{Z}_8$. Hence, $G/\text{Ker } f$ is a cyclic group of order 8. Thus, $G/\text{Ker } f$ has a normal subgroup H_1 of order 4 and a normal subgroup H_2 of order 2. By the correspondence theorem, there exist normal subgroups N_1 and N_2 of G such that Ker $f \subseteq N_1$, Ker $f \subseteq N_2$, $N_1/\text{Ker } f = H_1$, and $N_2/\text{Ker } f = H_2$. Thus,

$$8 = |G/\text{Ker } f| = [G : \text{Ker } f] = [G : N_1][N_1 : \text{Ker } f] = [G : N_1]4.$$

This implies that $[G : N_1] = 2$. Similarly, $[G : N_2] = 4$.

♢ **Exercise 3** Show that $4\mathbb{Z}/12\mathbb{Z} \simeq \mathbb{Z}_3$.

Solution: Define $f : 4\mathbb{Z} \to \mathbb{Z}_3$ by $f(4n) = [n]$ for all $4n \in 4\mathbb{Z}$. One can show that f is an epimorphism. Then from the first isomorphism theorem, $4\mathbb{Z}/\text{Ker } f \simeq \mathbb{Z}_3$. Now Ker $f = \{4n \in 4\mathbb{Z} \mid f(4n) = [0]\} = \{4n \in 4\mathbb{Z} \mid [n] = [0]\} = 12\mathbb{Z}$.

♢ **Exercise 4** Let G be a finite group and f be an automorphism of G such that for all $a \in G$, $f(a) = a$ if and only if $a = e$. Show that for all $g \in G$, there exists $a \in G$ such that $g = a^{-1}f(a)$.

Solution: Let $G = \{a_1, a_2, \ldots, a_n\}$. Let $S = \{a_1^{-1}f(a_1), \ldots, a_n^{-1}f(a_n)\}$. Then $S \subseteq G$. Next, we show that all elements of S are distinct. Now $a_i^{-1}f(a_i) = a_j^{-1}f(a_j)$ if and only if $f(a_i)f(a_j)^{-1} = a_i a_j^{-1}$ if and only if $f(a_i a_j^{-1}) = a_i a_j^{-1}$ if and only if $a_i a_j^{-1} = e$ if and only if $a_i = a_j$. This shows that all elements of S are distinct and so $|S| = n$. Thus, $S = G$. Let $g \in G$. Then $g \in S$. Hence, $g = a^{-1}f(a)$ for some $a \in G$.

♢ **Exercise 5** Let G be a finite group and f be an automorphism of G such that for all $a \in G$, $f(a) = a$ if and only if $a = e$. Suppose that $f^2 = i_G$, where i_G denotes the identity map. Prove that G is commutative.

Solution: Let $g \in G$. By Worked-Out Exercise 4, $g = a^{-1}f(a)$ for some $a \in G$. Then $g = i_G(g) = f^2(a^{-1}f(a)) = f(f(a^{-1}f(a))) = f(f(a^{-1})f^2(a)) = f(f(a)^{-1}a) = f(g^{-1})$. This implies that $f(g) = g^{-1}$ for all $g \in G$. Let $a, b \in G$. Then $(ab)^{-1} = f(ab) = f(a)f(b) = a^{-1}b^{-1} = (ba)^{-1}$ and so $ab = ba$. Hence, G is commutative.

♢ **Exercise 6** Let H be a subgroup of index 2 in a finite group G. If the order of H is odd and every element of $G \backslash H$ is of order 2, prove that H is commutative.

Solution: Since $[G : H] = 2$, H is a normal subgroup of G. Now $G = H \cup Hg$, where $g \notin H$. Then $\circ(g) = 2$. Define $f : G \to G$ by for all $a \in G$, $f(a) = gag^{-1}$. Then f is an automorphism of G. Now $f^2(a) = f(f(a)) = f(gag^{-1}) = g(gag^{-1})g^{-1} = g^2ag^{-2} = a$ since $g^2 = e$. Hence, $f^2 = i_G$. Since H is a normal subgroup of G, $f(h) = aha^{-1} \in H$ for all $h \in H$. Thus, f is also an automorphism of H. Let $h \in H$. Suppose $f(h) = h$. Then $ghg^{-1} = h$ or $gh = hg$. Since $gh \notin H$, $\circ(gh) = 2$. Therefore, $h^2 = g^2h^2 = (gh)^2 = e$. Since the order of H is odd, $h^2 = e$ implies that $h = e$. Hence, $f(h) = h$ if and only $h = e$. Thus, f is an automorphism of H such that $f^2 = i_G$ and $f(h) = h$ if and only if $h = e$. By Worked-Out Exercise 5, H is commutative.

♢ **Exercise 7** Show that $\text{Aut}(\mathbb{Z}_n) \simeq U_n$.

Solution: Define $\alpha : \text{Aut}(\mathbb{Z}_n) \to U_n$ by $\alpha(f) = f([1])$ for all $f \in \text{Aut}(\mathbb{Z}_n)$. Now $mf([1]) = f([m])$. Hence, $f([m]) = [0]$ if and only if m is divisible by n. Thus, $\circ(f([1])) = n$. This implies that $f([1]) \in U_n$ and so α is well defined. Let $f, g \in \text{Aut}(\mathbb{Z}_n)$. Then $\alpha(f \circ g) = (f \circ g)([1]) = f(g([1]))$. Suppose $g([1]) = [k]$. Then $\alpha(f \circ g) = f([k]) = kf([1]) = k[1]f([1]) = [k]f([1]) = f([1])g([1]) = \alpha(f)\alpha(g)$. Hence, α is a homomorphism. Now

$$\begin{aligned} \text{Ker } \alpha &= \{f \in \text{Aut}(\mathbb{Z}_n) \mid \alpha(f) = [1]\} \\ &= \{f \in \text{Aut}(\mathbb{Z}_n) \mid f([1]) = [1]\} \\ &= \{f \in \text{Aut}(\mathbb{Z}_n) \mid f \text{ is the identity map}\}. \end{aligned}$$

Hence, α is a monomorphism. Finally, we show that α is onto U_n. Let $[t] \in U_n$. Then t and n are relatively prime. Define $f : \mathbb{Z}_n \to \mathbb{Z}_n$ by $f([m]) = [mt]$ for all $[m] \in \mathbb{Z}_n$. Let $[r], [s] \in \mathbb{Z}_n$. Suppose $[r] = [s]$. Then $r - s = nq$ for some $q \in \mathbb{Z}$. Thus, $rt - st = nqt$. Hence, $[rt] = [st]$, proving that f is well defined. Clearly f is a homomorphism. Suppose $f([r]) = f([s])$. Then $[rt] = [st]$ and so n divides $rt - st = (r - s)t$. Since t and n are relatively prime, n divides $r - s$. Therefore, $[r] = [s]$. This implies that f is one-one. Since Z_n is finite therefore we find that f is onto. Hence, $f \in \text{Aut}(\mathbb{Z}_n)$. Now $\alpha(f) = f([1]) = [t]$ shows that α is onto U_n. Thus, α is an isomorphism. Consequently, $\text{Aut}(\mathbb{Z}_n) \simeq U_n$.

Exercises

1. Let $\mathbb{R}^*$ be the multiplicative group of all nonzero real numbers and $T = \{1, -1\}$. Then T is a subgroup of $\mathbb{R}^*$. Prove that the quotient group $\mathbb{R}^*/T$ is isomorphic to the multiplicative group $\mathbb{R}^+$ of positive real numbers.
2. For any positive integer n, prove that $\mathbb{Z}/n\mathbb{Z} \simeq \mathbb{Z}_n$.

3. Show that $8\mathbb{Z}/56\mathbb{Z} \simeq \mathbb{Z}_7$ and $4\mathbb{Z}/8\mathbb{Z} \simeq \mathbb{Z}_2$
4. For any two positive integers m and n prove that $m\mathbb{Z}/mn\mathbb{Z} \simeq \mathbb{Z}_n$.
5. Let G be a group and A and B be normal subgroups of G such that $A \simeq B$. Show by an example that $G/A \not\simeq G/B$.
6. Let G be the group of symmetries of the square and K_4 the Klein 4-group.

   Show that the mapping $f : G \to K_4$ defines a homomorphism of G onto K_4, where $f(r_{180}) = f(r_{360}) = e$, $f(r_{90}) = f(r_{270}) = a$, $f(h) = f(v) = b$, $f(d_1) = f(d_2) = c$.
7. In Exercise 6, exhibit the one-one inclusion preserving correspondence between the subgroups of G containing $Z(G)$ and the subgroups of K_4.
8. Let G and K_4 be as in Exercise 6. Let g be the natural homomorphism of G onto $G/Z(G)$, where $Z(G)$ is the center of G. Prove that $Z(G) = \text{Ker } f$ and exhibit the isomorphism h of $G/Z(G)$ onto K_4 such that $f = h \circ g$.
9. Show that $\mathbb{Z}_8$ is not a homomorphic image of $\mathbb{Z}_{15}$.
10. Show that $\mathbb{Z}_9$ is not a homomorphic image of $\mathbb{Z}_3 \times \mathbb{Z}_3$.
11. Show that if there exists an epimorphism from a finite group G onto the group $\mathbb{Z}_{15}$, then G has normal subgroups of indices 5 and 3, respectively.
12. Partition the following collection of groups into subcollections of groups such that any two groups in the same subcollection are isomorphic.

    (i) $(\mathbb{Z}, +)$, (ii) $(\mathbb{Z}_6, +)$, (iii) $(\mathbb{Z}_2, +)$, (iv) S_2, (v) S_6, (vi) $(17\mathbb{Z}, +)$, (vii) $(3\mathbb{Z}, +)$, (vii) $(\mathbb{Q}, +)$, (ix) $(\mathbb{R}, +)$, (x)$(\mathbb{R}^*, \cdot)$, (xi) $(\mathbb{R}^+, \cdot)$, (xii) $(\mathbb{Q}^*, \cdot)$, (xiii) $(\mathbb{C}^*, \cdot)$, (xiv) $(\langle\pi\rangle, \cdot)$, where $\mathbb{R}^*$ denotes the set of nonzero real numbers, $\mathbb{Q}^*$ denotes the set of nonzero rational numbers, $\mathbb{C}^*$ denotes the set of nonzero real numbers, $\mathbb{R}^+$ denotes the set of positive real numbers, and $(\langle\pi\rangle, \cdot)$ is the cyclic subgroup of $(\mathbb{R}^+, \cdot)$ generated by π.
13. Show that

    (a) $\text{Aut}(\mathbb{Z}_5) \simeq \mathbb{Z}_4$.

    (b) $\text{Aut}(\mathbb{Z}_8) \simeq$ Klein 4-group.

    (c) $\text{Aut}(\mathbb{Z}_2 \times \mathbb{Z}_2) \simeq S_3$

    (d) $\text{Aut}(\mathbb{Z}) \simeq Z_2$

    (e) $\text{Aut}(\mathbb{Q}, +)$ contains infinite number of elements.
14. Find all automorphisms of the group $\mathbb{Z}_6$.
15. Show that $|\text{Aut}(\mathbb{Z}_p)| = p - 1$, where p is a prime.
16. Prove that $\text{Inn}(S_3) \simeq S_3 \simeq \text{Aut}(S_3)$.
17. Determine $\text{Aut}(S_4)$.
18. Let G be a cyclic group of order n and ϕ be the Euler ϕ-function. Prove that $|\text{Aut}(G)| = \phi(n)$.
19. Let G be a group such that $Z(G) = \{e\}$. Prove that $Z(\text{Aut}(G)) = \{e\}$.
20. Let G be a group and H be a subgroup of G. H is called a **characteristic** subgroup of G if $f(H) \subseteq H$ for all $f \in \text{Aut}(G)$.

    (a) Show that every characteristic subgroup of G is a normal subgroup of G.

    (b) Give an example of a group G and a subgroup H such that H is a normal subgroup of G, but H is not a characteristic subgroup of G.

    (c) Show that $Z(G)$ is a characteristic subgroup of G.

    (d) Let H and K be characteristic subgroups of G. Show that HK and $H \cap K$ are characteristic subgroups of G.

    (e) Let H and K be subgroups of G such that $H \subseteq K$. Show that if K is a normal subgroup of G and H is a characteristic subgroup of G, then H is a normal subgroup of G.

    (f) Let H and K be subgroups of G such that $H \subseteq K$. Show that if H is a characteristic subgroup of K and K is a characteristic subgroup of G, then H is a characteristic subgroup of G.

    (g) Suppose G is cyclic. Show that every subgroup of G is a characteristic subgroup of G.

21. Show that the only characteristic subgroups of $(\mathbb{Q}, +)$ are $\{0\}$ and $\mathbb{Q}$.
22. Which of the following statements are true? Justify.
    (a) Any epimorphism of $\mathbb{Z}$ onto $\mathbb{Z}$ is an isomorphism.
    (b) Any epimorphism of a group G onto G is an isomorphism.
    (c) The quotient group $4\mathbb{Z}/64\mathbb{Z}$ has five subgroups.
    (d) $\mathbb{Z}_5$ has five homomorphic images.
    (e) $2\mathbb{Z}/6\mathbb{Z}$ is a subgroup of $\mathbb{Z}/6\mathbb{Z}$.
    (f) There exist four subgroups of $\mathbb{Z}$ which contain $10\mathbb{Z}$ as a subgroup.
    (g) Let G and H be two groups, A be a normal subgroup of G, and B be a normal subgroup of H. If $G \simeq H$ and $A \simeq B$, then $G/A \simeq H/B$.

5.3 The Groups D_4 and Q_8

In Section 5.1, we saw that there are two types of groups of order 4 and two types of groups of order 6. In this section, we wish to classify all noncommutative groups of order 8. We will consider finite commutative groups in Chapter 9. First we introduce two groups D_4 and Q_8 and study these groups in detail. The study of these groups will eventually lead us to the classification of noncommutative groups of order 8.

Definition 5.3.1 *A group G is called a **dihedral** group of degree 4 if G is generated by two elements a and b satisfying the relations*

$$\circ(a) = 4, \ \circ(b) = 2, \text{ and } ba = a^3b.$$

Example 5.3.2 *Let G be the subgroup of GL(2,R)(Example 2.1.10) generated by the matrices*

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

Then $\circ(A) = 4$ and $\circ(B) = 2$. Now

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

and

$$A^3B = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus, $BA = A^3B$. Hence, G is a dihedral group of degree 4.

Example 5.3.3 *Consider S_4. Let G be the subgroup of S_4 such that G is generated by the permutations*

$$a = (1\ 2\ 3\ 4) \text{ and } b = (2\ 4).$$

Then $a^2 = (1\ 3) \circ (2\ 4)$, $a^3 = (1\ 4\ 3\ 2)$, $a^4 = e$, $b^2 = e$, and $b \circ a = (1\ 4) \circ (2\ 3) = a^3 \circ b$. Hence, $\circ(a) = 4$, $\circ(b) = 2$, and $b \circ a = a^3 \circ b$. Thus, G is a dihedral group of degree 4.

The following theorem reveals some interesting properties of D_4. These properties are similar to the properties listed in Example 4.1.21 for D_3.

Theorem 5.3.4 *Let G be a dihedral group of degree 4 generated by the elements a and b such that*

$$\circ(a) = 4, \ \circ(b) = 2, \text{ and } ba = a^3b.$$

Then the following assertions hold.
(i) Every element of G is of the form a^ib^j, $0 \leq i < 4$, $0 \leq j < 2$.
(ii) G has exactly eight elements, i.e., $|G| = 8$.
(iii) G is a noncommutative group.

Proof. (i) Since $G = \langle a, b\rangle$,

$$G = \{a^{i_1}b^{j_1}a^{i_2}b^{j_2}\cdots a^{i_n}b^{j_n} \mid i_t, j_t \in \mathbb{Z}, 1 \le t \le n,\ n \in \mathbb{N}\}.$$

Since $a^{-1} = a^3$ and $b^{-1} = b$,every element of G can be expressed in the form $a^{i_1}b^{j_1}a^{i_2}b^{j_2}\cdots a^{i_n}b^{j_n}$ where $i_t \ge 0$ and $j_t \ge 0$. Again since $ba = a^3b$ it follows that every element of G is of the form a^nb^m, where n, m are nonnegative integers. Now $a^4 = e$, $b^2 = e$. These imply that every element of G is of the form a^ib^j, $0 \le i < 4$, $0 \le j < 2$.

(ii) By (i), every element of G is of the form a^ib^j, $0 \le i < 4$, $0 \le j < 2$. Thus, $|G| \le 8$. Since $\circ(a) = 4$, it follows that e, a, a^2, a^3 are distinct elements of G. Then b, ab, a^2b, a^3b are also distinct elements of G. Also, since $a^{-1} = a^3$, $b^{-1} = b$, and $a \ne b \ne e$,

$$\{e, a, a^2, a^3\} \cap \{b, ab, a^2b, a^3b\} = \emptyset.$$

Thus, $G = \{e, a, a^2, a^3, b, ab, a^2b, a^3b\}$. Hence, G has eight elements.

(iii) Suppose $ab = ba$. Then $ab = a^3b$. This implies that $a^2 = e$, which is a contradiction. Hence, $ab \ne ba$, proving that G is noncommutative. ■

It is easy to see that any two dihedral groups of degree 4 are isomorphic. Hence, there exists only one dihedral group (up to isomorphism) of degree 4. We denote a dihedral group of degree 4 by D_4.

We now describe all subgroups of D_4.

In D_4,

$$\circ(a) = 4, \circ(a^2) = 2, \circ(a^3) = 4, \circ(b) = 2,$$
$$(ab)^2 = abab = aa^3bb = e,$$
$$(a^2b)^2 = a^2ba^2b = a^2(a^3b)ab = abab = e,$$
$$(a^3b)^2 = a^3ba^3b = a^3(a^3b)a^2b = a^2ba^2b = e.$$

From this, it follows that $H_1 = \{e, a^2\}$, $H_2 = \{e, b\}$, $H_3 = \{e, ab\}$, $H_4 = \{e, a^2b\}$, and $H_5 = \{e, a^3b\}$ are subgroups of order 2. By Lagrange's theorem, D_4 has no subgroups of order 3, 5, 6, or 7. Now

$$T_1 = \{e, a, a^2, a^3\}$$
$$T_2 = \{e, a^2, b, a^2b\}$$
$$T_3 = \{e, ab, a^2, a^3b\}$$

are subgroups of order 4. We ask the reader to verify that $\{e\}$, H_1, H_2, H_3, H_4, H_5, T_1, T_2, T_3, and D_4 are the only subgroups of D_4.

It is interesting to note in D_4 that H_5 is a normal subgroup of T_3 and T_3 is a normal subgroup of D_4, but H_5 is not a normal subgroup of D_4. We also note that every nontrivial subgroup of D_4 is of order 2 or 4. Therefore, every nontrivial subgroup of D_4 is commutative. However, since T_2 is a nontrivial subgroup of D_4 and T_2 is not cyclic, it follows that not every nontrivial subgroup of D_4 is cyclic. Finally, we also note that D_4 is isomorphic to Sym, the group of symmetries of a square (page 47). This follows from Theorem 5.3.4 and the group table of the group of symmetries of the square given on page 49.

Next, we consider Q_8.

Definition 5.3.5 *A group G is called a **quaternion** group if G is generated by two elements a, b satisfying the relation*

$$\circ(a) = 4,\ a^2 = b^2,\ \text{and } ba = a^3b.$$

Example 5.3.6 *Let T be the group of all 2×2 invertible matrices over $\mathbb{C}$ under usual matrix multiplication. Let G be the subgroup of T generated by the matrices*

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \text{ and } B = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.$$

Then $\circ(A) = 4$ and

$$A^2 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = B^2.$$

Now

$$BA = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix}$$

and

$$A^3B = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix} = \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix}.$$

Thus, $BA = A^3B$. Hence, G is a quaternion group.

We leave the proof of the following theorem, which is similar to the proof of Theorem 5.3.4, as an exercise.

Theorem 5.3.7 *Let G be a quaternion group generated by the elements a and b such that*

$$\circ(a) = 4,\ a^2 = b^2,\ \text{and } ba = a^3b.$$

Then the following assertions hold.
(i) Every element of G is of the form a^ib^j, $0 \leq i < 4$, $0 \leq j < 2$.
(ii) G has exactly eight elements, i.e., $|G| = 8$.
(iii) G is a noncommutative group. ■

It is easy to see that any two quaternion groups are isomorphic. Hence, there exists only one quaternion group (up to isomorphism) and we denote it by Q_8.

Next, we determine all subgroups of Q_8.

Let $Q_8 = \langle a, b\rangle$, where $\circ(a) = 4$, $a^2 = b^2$, and $ba = a^3b$. Then

$$Q_8 = \{e, a, a^2, a^3, b, ab, a^2b, a^3b\}.$$

In Q_8,

$$\circ(a) = 4, \circ(a^2) = 2, \circ(a^3) = 4, \circ(b) = 4.$$

Now

$$(ab)^2 = abab = aa^3bb = b^2 = a^2.$$

Thus, $\circ(ab) = 4$. Also,

$$(a^2b)^2 = a^2ba^2b = a^2(a^3b)ab = a^5bab = abab$$

and

$$(a^3b)^2 = a^3ba^3b = a^3(a^3b)a^2b = a^2ba^2b.$$

Hence, $\circ(a^2b) = 4$ and $\circ(a^3b) = 4$. It now follows that $H_0 = \{e\}$, $H_1 = \{e, a^2\}$, $H_2 = \{e, a, a^2, a^3\}$, $H_3 = \{e, ab, a^2, a^3b\}$, and $H_4 = \{e, b, a^2, a^2b\}$ are subgroups of Q_8. We ask the reader to verify that H_0, H_1, H_2, H_3, H_4, and Q_8 are the only subgroups of Q_8.

Since $[Q_8 : H_2] = [Q_8 : H_3] = [Q_8 : H_4] = 2$, H_2, H_3, and H_4 are normal subgroups of Q_8. Now $ba^2b^{-1} = baab^{-1} = a^3bab^{-1} = a^3a^3bb^{-1} = a^2 \in H_1$. Since $Q_8 = \langle a, b\rangle$, H_1 is a normal subgroup of Q_8. Thus, every subgroup of Q_8 is a normal subgroup of G. It is also interesting to observe that all proper subgroups of Q_8 are cyclic.

Theorem 5.3.8 $D_4 \not\simeq Q_8$.

Proof. We note from the above discussion that Q_8 contains six elements of order 4 while D_4 contains only two elements of order 4. Hence, $D_4 \not\simeq Q_8$. ■

The next theorem classifies all noncommutative groups of order 8.

Theorem 5.3.9 *There exist (up to isomorphism) only two noncommutative nonisomorphic groups of order* 8.

Proof. Let G be a noncommutative group of order 8. Since $|G|$ is even, there exists an element $u \in G$, $u \neq e$, such that $u^2 = e$. If $x^2 = e$ for all $x \in G$, then G is commutative, a contradiction. Thus, there exists $a \in G$ such that $a^2 \neq e$. Since $\circ(a) \mid 8$, $\circ(a) = 4$ or 8. If $\circ(a) = 8$, then G is cyclic and hence commutative, a contradiction. Thus, $\circ(a) = 4$. Let $H = \{e, a, a^2, a^3\}$. Then H is a subgroup of G of index 2 and so H is a normal subgroup of G. Let $b \in G$ be such that $b \notin H$. Then $G = H \cup Hb$ and $H \cap Hb = \emptyset$. This implies that

$$G = \{e, a, a^2, a^3, b, ab, a^2b, a^3b\} = \langle a, b\rangle.$$

Now $bab^{-1} \in H$. If $bab^{-1} = e$, then $a = e$, a contradiction. Thus, $bab^{-1} \neq e$. If $bab^{-1} = a$, then $ab = ba$ and hence G is commutative, a contradiction. If $bab^{-1} = a^2$, then $ba^2b^{-1} = (bab^{-1})^2 = a^4 = e$ and so $a^2 = e$, a contradiction. Therefore, $bab^{-1} = a^3$ and so $ba = a^3b$. Since $|G/H| = 2$ and $b \notin H$, $\circ(Hb) = 2$. Hence, $b^2 \in H$. If $b^2 = a$ or a^3, then $\circ(b) = 8$ and so G is commutative, a contradiction. Therefore, either $b^2 = e$ or $b^2 = a^2$. It now follows that if G is a noncommutative group of order 8, then either

$$G = \langle a, b\rangle \text{ such that } \circ(a) = 4,\ \circ(b) = 2 \text{ ,and } ba = a^3b$$

or

$$G = \langle a, b\rangle \text{ such that } \circ(a) = 4,\ b^2 = a^2, \text{ and } ba = a^3b.$$

In the first case, $G \simeq D_4$ and in the second case, $G \simeq Q_8$. ■

Worked-Out Exercises

♢ **Exercise 1** Find $Z(D_4)$.

Solution: It is known that $Z(D_4)$ is a normal subgroup of D_4. Now D_4 has six normal subgroups: D_4, $\{e\}$, $H_1 = \{e, a^2\}$, $T_1 = \{e, a, a^2, a^3\}$, $T_2 = \{e, a^2, b, a^2b\}$, $T_3 = \{e, ab, a^2, a^3b\}$. Since $ab \neq ba$, D_4, T_1, and T_2 cannot be $Z(D_4)$. If $(ab)b = b(ab)$, then $a = (ba)b = a^3b^2 = a^3$ and so $a^2 = e$, a contradiction. Hence, $T_3 \neq Z(D_4)$. Now $a^2b = a^6b = a^3(a^3b) = a^3(ba) = (ba)a = ba^2$. Hence, $a^2 \in Z(D_4)$. Thus, $Z(D_4) = \{e, a^2\} = H_1$.

♢ **Exercise 2** Prove that $D_4/Z(D_4)$ is isomorphic to K_4 and hence find $\mathrm{Inn}(D_4)$.

Solution: By Corollary 5.2.18, $\mathrm{Inn}(D_4) \simeq D_4/Z(D_4)$. Now $D_4/Z(D_4)$ is a group of order 4 and

$$D_4/Z(D_4) = \{eZ(D_4),\ aZ(D_4),\ bZ(D_4),\ abZ(D_4)\}.$$

Since $a^2 \in Z(D_4)$, $b^2 = e$, and $(ab)^2 = e$, we find that each nonidentity element of $D_4/Z(D_4)$ is of order 2. Hence, $D_4/Z(D_4) \simeq K_4$, the Klein 4-group.

Exercises

1. In D_4, find subgroups H and K such that K is a normal subgroup of H and H is a normal subgroup of D_4, but K is not a normal subgroup of D_4.
2. Show that Q_8 is the union of three subgroups each of index 2.
3. Find all homomorphic images of D_4.
4. Find all homomorphic images of Q_8.

Group Actions

As previously mentioned, the theory of groups first dealt with permutation groups. Later the notion of an abstract group was introduced in order to examine properties of permutation groups which did not refer to the set on which the permutations acted. However, one is primarily interested in permutation groups in geometry. Also, permutation groups are used in counting techniques that are important in finite group theory. An example of this can be seen in the proof of Lagrange's theorem. We extend the notion of a permutation on a set to a group action on a set. We use the notion of a group action on a set to determine, via counting techniques, important properties of finite groups.

Let G be a group and S a nonempty set. A **(left) action** of G on S is a function $\cdot : G \times S \to S$ (usually denoted by $\cdot(g, x) \to g \cdot x$) such that
(i) $(g_1g_2) \cdot x = g_1 \cdot (g_2 \cdot x)$, and
(ii) $e \cdot x = x$, where e is the identity of G
for all $x \in S$, $g_1, g_2 \in G$.

Note: If no confusion arises, we write gx for $g \cdot x$.

If there is a left action of G on S, we say that G acts on S on the left and S is a **G-set.**

Example 5.3.10 *Let G be a permutation group on a set S. Define a left action of G on S by*

$$\sigma x = \sigma(x)$$

for all $\sigma \in G$, $x \in S$. Let $x \in S$. Now $ex = e(x) = x$, where e is the identity permutation on S. Let $\sigma_1, \sigma_2 \in G$. Then $(\sigma_1 \circ \sigma_2) \cdot x = (\sigma_1 \circ \sigma_2)(x) = \sigma_1(\sigma_2(x)) = \sigma_1 \cdot (\sigma_2(x)) = \sigma_1 \cdot (\sigma_2 \cdot x)$. Hence, S is a G-set.

Example 5.3.11 *Let G be a group and H be a normal subgroup of G. Define a left action of G on H by*

$$(g, h) \to ghg^{-1}$$

for all $g \in G$, $h \in H$. We denote this by $g \cdot h = ghg^{-1}$. Let $h \in H$. Now $e \cdot h = ehe^{-1} = ehe = h$. Let $g_1, g_2 \in G$. Then $(g_1g_2) \cdot h = (g_1g_2)h(g_1g_2)^{-1} = (g_1g_2)h(g_2^{-1}g_1^{-1}) = g_1(g_2hg_2^{-1})g_1^{-1} = g_1(g_2 \cdot h)g_1^{-1} = g_1 \cdot (g_2 \cdot h)$. Hence, H is a G-set.

Theorem 5.3.12 *Let S be a G-set, where G is a group and S is a nonempty set. Define a relation $\sim$ on S by for all $a, b \in S$,*

$$a \sim b \text{ if and only if } ga = b \text{ for some } g \in G.$$

Then $\sim$ is an equivalence relation on S.

Proof. Since for all $a \in S$, $ea = a$, $a \sim a$ for all $a \in S$. Thus, $\sim$ is reflexive. Let $a, b, c \in S$. Suppose $a \sim b$. Then $ga = b$ for some $g \in G$, which implies that $g^{-1}b = g^{-1}(ga) = (g^{-1}g)a = ea = a$. Hence, $b \sim a$ and so $\sim$ is symmetric. Now suppose $a \sim b$ and $b \sim c$. Then there exist g_1, $g_2 \in G$ such that $g_1a = b$ and $g_2b = c$. Thus, $(g_2g_1)a = g_2(g_1a) = g_2b = c$ and so $a \sim c$. Hence, $\sim$ is transitive. Consequently, $\sim$ is an equivalence relation. ■

Definition 5.3.13 *Let S be a G-set, where G is a group and S is a nonempty set. The equivalence classes determined by the equivalence relation of Theorem* 5.3.12 *are called the* ***orbits*** *of G on S.*

For $a \in S$, the orbit containing a is denoted by $[a]$.

Lemma 5.3.14 *Let G be a group and S be a G-set. For all $a \in S$, the subset*

$$G_a = \{g \in G \mid ga = a\}$$

is a subgroup of G.

Proof. Let $a \in S$. Since $ea = a$, $e \in G_a$ and so $G_a \neq \emptyset$. Let $g, h \in G_a$. Then $ga = a$ and $ha = a$. This implies that $(gh)a = g(ha) = ga = a$ and so $gh \in G_a$. Now $h^{-1}a = h^{-1}(ha) = (h^{-1}h)a = ea = a$. Thus, $h^{-1} \in G_a$. Hence, G_a is a subgroup of G. ■

The subgroup G_a of Lemma 5.3.14 is called the **stabilizer** of a or the **isotropy group** of a.

Lemma 5.3.15 *Let G be a group and S be a G-set. For all $a \in S$,*

$$[G : G_a] = |[a]|.$$

Proof. Let $a \in S$. Let $\mathcal{L}$ be the set of all left cosets of G_a in G. Now

$$[a] = \{b \in S \mid a \sim b\} = \{b \in S \mid ga = b \text{ for some } g \in G\} = \{ga \mid g \in G\}.$$

We now show that there exists a one-one function from $\mathcal{L}$ onto $[a]$. Define

$$f : \mathcal{L} \to [a]$$

by

$$f(gG_a) = ga$$

for all $gG_a \in \mathcal{L}$. Let g_1, $g_2 \in G$. Then $g_1G_a = g_2G_a$ if and only if $g_2^{-1}g_1 \in G_a$ if and only if $g_2^{-1}(g_1a) = (g_2^{-1}g_1)a = a$ if and only if $g_1a = g_2a$. Thus, f is a one-one function from $\mathcal{L}$ into $[a]$. Let $b \in [a]$. Then there exists $g \in G$ such that $ga = b$. Thus, $f(gG_a) = ga = b$. This implies that f is onto $[a]$. Consequently, $[G : G_a] = |\mathcal{L}| = |[a]|$. ■

Theorem 5.3.16 *Let G be a group and S be a G-set. If S is finite, then*

$$|S| = \sum_{a \in A} [G : G_a],$$

where A is a subset of S containing exactly one element from each orbit $[a]$.

Proof. By Theorem 5.3.12, S can be partitioned as the union of orbits. Therefore,

$$S = \cup_{a \in A}[a].$$

Hence,

$$|S| = \sum_{a \in A} |[a]| = \sum_{a \in A} [G : G_a] \text{ by Lemma 5.3.15.}$$

■

Theorem 5.3.17 *Let S be a finite G-set, where G is a group of order p^n (p a prime). Let $S_0 = \{a \in S \mid ga = a \text{ for all } g \in G\}$. Then*

$$|S| \equiv_p |S_0|.$$

Proof. By Theorem 5.3.16,

$$|S| = \sum_{a \in A} [G : G_a],$$

where A is a subset of S containing exactly one element from each orbit $[a]$ of G. Now $a \in S_0$ if and only if $ga = a$ for all $g \in G$, i.e., if and only if $[a] = \{a\}$. Hence,

$$|S| = |S_0| + \sum_{a \in A \backslash S_0} \frac{|G|}{|G_a|}.$$

Since $|G_a| \neq |G|$ for all $a \in A \backslash S_0$, $\frac{|G|}{|G_a|}$ is some power of p for all $a \in A \backslash S_0$. Thus, $\frac{|G|}{|G_a|}$ is divisible by p, proving that $|S| \equiv_p |S_0|$ ■

Corollary 5.3.18 *Let G be a finite group and H be a subgroup of G such that $|H| = p^k$, where p is a prime and k is a nonnegative integer. Then*

$$[G : H] \equiv_p [N(H) : H],$$

where $N(H) = \{g \in G \mid gHg^{-1} = H\}$.

Proof. Let $S = \{xH \mid x \in G\}$. Define a left action of H on S by $h(xH) = (hx)H$ for all $h \in H$, $xH \in S$. Then S is an H-set. Let $S_0 = \{xH \in S \mid h(xH) = xH \text{ for all } h \in H\}$. By the above theorem, $|S| \equiv_p |S_0|$. Now $xH \in S_0$ if and only if $h(xH) = xH$ for all $h \in H$ if and only if $x^{-1}hx \in H$ for all $h \in H$ if and only if $x^{-1}Hx \subseteq H$. Now $\left|x^{-1}Hx\right| = |H|$. Hence, $xH \in S_0$ if and only if $x^{-1}Hx \subseteq H$ if and only if $x^{-1}Hx = H$ (since H is finite and $\left|x^{-1}Hx\right| = |H|$) if and only if $x \in N(H)$. This shows that S_0 is the set of all left cosets of H in $N(H)$. Thus, $|S_0| = [N(H) : H]$. Also, $|S| = [G : H]$. Hence, $[G : H] \equiv_p [N(H) : H]$. ■

Theorem 5.3.19 *Let G be a group and S be a G-set. Then the left action of G on S induces a homomorphism from G into $A(S)$, where $A(S)$ is the group of all permutations of S.*

Proof. Let $g \in G$. Define $\tau_g : S \to S$ by $\tau_g(a) = ga$ for all $a \in S$. Let $a, b \in S$. Then $\tau_g(a) = \tau_g(b)$ if and only if $ga = gb$ if and only if $a = b$. Therefore, τ_g is a one-one function. Now $b = g(g^{-1}b) = \tau_g(g^{-1}b)$ and $g^{-1}b \in S$. This shows that τ_g is onto S. Thus, $\tau_g \in A(S)$. Let g_1, $g_2 \in G$. Then $\tau_{g_1g_2}(a) = (g_1g_2)a = g_1(g_2a) = \tau_{g_1}(g_2a) = \tau_{g_1}(\tau_{g_2}(a)) = (\tau_{g_1} \circ \tau_{g_2})(a)$ for all $a \in S$. This implies that $\tau_{g_1g_2} = \tau_{g_1} \circ \tau_{g_2}$. Define

$$\psi : G \to A(S)$$

by

$$\psi(g) = \tau_g$$

for all $g \in G$. Then ψ is a function. Now $\psi(g_1g_2) = \tau_{g_1g_2} = \tau_{g_1} \circ \tau_{g_2} = \psi(g_1) \circ \psi(g_2)$ for all g_1, $g_2 \in G$. This proves that ψ is a homomorphism.
■

The following corollary, which is known as the Extended Cayley's theorem, follows from the above theorem.

Theorem 5.3.20 *Extended Cayley's theorem: Let G be a group and H be a subgroup of G. Let $S = \{aH \mid a \in G\}$. Then there exists a homomorphism ψ from G into $A(S)$ (the group of all permutations on S) such that Ker $\psi \subseteq H$.*

Proof. First we note that S is a G-set, where the left action of G on S is defined by $g(aH) = (ga)H$ for all $g \in G$. This left action induces the homomorphism ψ of Theorem 5.3.19. Now

$$\begin{aligned} \text{Ker } \psi &= \{g \in G \mid \psi(g) = \tau_g = \text{ the identity mapping on } S\} \\ &= \{g \in G \mid \tau_g(aH) = aH \text{ for all } aH \in S\} \\ &= \{g \in G \mid g(aH) = aH \text{ for all } aH \in S\}. \end{aligned}$$

Let $g \in \text{Ker } \psi$. Then $g(aH) = aH$ for all $aH \in S$. In particular, $gH = H$. Thus, $g \in H$. Hence, Ker $\psi \subseteq H$. ■

Corollary 5.3.21 *Let G be a group and H be a subgroup of G of index n .Then there exists a homomorphism ψ from G into S_n such that Ker $\psi \subseteq H$.*

Proof. Because $[G : H] = n$, the group $A(S)$ of all permutations on S is isomrphic to S_n. Hence the corollary follows from the theorem. ■

Corollary 5.3.22 *Let H be a subgroup of a group G of index a prime integer p. Then H is isomorphic to a subgroup of S_p.*

Corollary 5.3.23 *Let G be a finite group and H be a proper subgroup of G of index n such that $|G|$ does not divide $n!$ Then G contains a nontrivial normal subgroup.*

Proof. From Corollary 5.3.20, Ker $\psi \subseteq H$ and $G/\text{Ker } \psi$ is isomorphic to a subgroup of S_n, where ψ is as defined in Corollary 5.3.20. Therefore, $|G/\text{Ker } \psi|$ divides $n!$ But $|G|$ does not divide $n!$ Hence, $|\text{Ker } \psi| \neq 1$, proving that Ker ψ is a nontrivial normal subgroup of G. ■

Definition 5.3.24 *Let G be a group and S be a G-set. Let $a \in S$, $g \in G$. Then a is called **fixed** by g if $ga = a$. If $ga = a$ for all $g \in G$, then a is called fixed by G.*

Theorem 5.3.25 (Burnside) *Let S be a finite nonempty set and G be a finite group. If S is a G-set, then the number of orbits of G is*

$$\frac{1}{|G|}\sum_{g\in G} F(g),$$

where $F(g)$ is the number of elements of S fixed by g.

Proof. Let $T = \{(g, a) \in G \times S \mid ga = a\}$. Since $F(g)$ is the number of elements $a \in S$ such that $(g, a) \in T$, it follows that $|T| = \sum_{g\in G} F(g)$. Also, $|G_a|$ is the number of elements $g \in G$ such that $(g, a) \in T$. Hence, $|T| = \sum_{a\in S} |G_a|$.

Let $S = [a_1] \cup [a_2] \cup \cdots \cup [a_k]$, where $\{[a_1], [a_2], \ldots, [a_k]\}$ is the set of all distinct orbits of G on S. Then

$$\sum_{g\in G} F(g) = \sum_{a\in[a_1]} |G_a| + \sum_{a\in[a_2]} |G_a| + \cdots + \sum_{a\in[a_k]} |G_a|.$$

Suppose a, b are in the same orbit. Then $[a] = [b]$ and $[G : G_a] = |[a]| = |[b]| = [G : G_b]$. This implies

$$\frac{|G|}{|G_a|} = \frac{|G|}{|G_b|}$$

and so $|G_a| = |G_b|$. Thus,

$$\begin{aligned} \textstyle\sum_{g\in G} F(g) &= |[a_1]|\,|G_{a_1}| + |[a_2]|\,|G_{a_2}| + \cdots + |[a_k]|\,|G_{a_k}| \\ &= \tfrac{|G|}{|G_{a_1}|}\,|G_{a_1}| + \tfrac{|G|}{|G_{a_2}|}\,|G_{a_2}| + \cdots + \tfrac{|G|}{|G_{a_k}|}\,|G_{a_k}| \\ &= k\,|G|, \end{aligned}$$

where k is the number of distinct orbits. Consequently,

$$k = \frac{1}{|G|}\sum_{g\in G} F(g).$$

■

Worked-Out Exercises

♢ **Exercise 1** Let S be a finite G-set, where G is a group of order p^n (p a prime) such that p does not divide $|S|$. Show that there exists

Solution: Let $S_0 = \{a \in S \mid ga = a \text{ for all } g \in G\}$. By Worked-Out Exercise 1, $|S| \equiv_p |S_0|$. Since p does not divide $|S|$, p does not divide $|S_0|$. Thus, $|S_0| \neq 0$. This shows that there exists $a \in S_0$. Thus, a is fixed by G.

◇ **Exercise 2** Let G be a finite group. Let H be a subgroup of G of index p, where p is the smallest prime dividing the order of G. Show that H is a normal subgroup of G.

Solution: Let $S = \{aH \mid a \in G\}$. Since $[G : H] = p$, by Extended Cayley's theorem there exista a homomorphism $\psi : G \to A(S)$ such that Ker $\psi \subseteq H$. Now $G/\text{Ker } \psi$ is isomorphic to a subgroup of $A(S)$. Therefore, $|G/\text{Ker } \psi|$ divides $|A(S)| = p!$ Let $|G/\text{Ker } \psi| = n$. Then $n = [G : H][H : \text{Ker } \psi] \geq p$. Let $n = p_1p_2 \cdots p_k$, where p_i are prime integers, $i = 1, 2, \ldots, k$. Since p_i divides $|G|$ and p is the smallest prime dividing the order of G, $p_i \geq p$ for all $i = 1, 2, \ldots, k$. Since n divides $p!$, we have each p_i divides $p!$. Since each p_i is a prime and $p_i \geq p$, we must have $i = 1$ and $p_i = p$. Thus, $n = p$. This implies that $[H : \text{Ker } \psi] = 1$. Hence, $H = \text{Ker } \psi$ and so H is a normal subgroup of G.

♢ **Exercise 3** Let G be a group of order pn, p a prime, and $p \geq n$. If H is a subgroup of order p in G, prove that H is a normal subgroup of G.

Solution: Let $S = \{aH \mid a \in G\}$. Now $|S| = [G : H] = \frac{|G|}{|H|} = \frac{pn}{p} = n$. By Extended Cayley's theorem there exista a homomorphism $\psi : G \to A(S)$ such that Ker $\psi \subseteq H$. Since $|H| = p$, either Ker $\psi = \{e\}$ or Ker $\psi = H$. If Ker $\psi = \{e\}$, then G is isomorphic to a subgroup of $A(S)$. This implies that $|G|$ divides $|A(S)|$, i.e., $pn \mid n!$ Therefore, $p \mid (n-1)!$ Since $p \geq n$,p does not divide $(n-1)!$ Thus, Ker $\psi = H$. Hence, H is a normal subgroup of G.

♢ **Exercise 4** Let G be a group. Show that G is isomorphic to a subgroup of $A(G)$. (This is Cayley's theorem. Here we want to prove this result by the group action method.)

Solution: G is a G-set, where the left action of G on G is defined by the group operation. This left action induces a homomorphism $\psi : G \to A(G)$ defined by $\psi(g) = \tau_g$, where $\tau_g(a) = ga$ for all $a, g \in G$. Now Ker $\psi = \{g \in G \mid \tau_g =$ identity permutation on $G\} = \{g \in G \mid ga = a$ for all $a \in G\} = \{e\}$. Hence, ψ is a monomorphism.

Exercises

1. Show that $I_3 = \{1, 2, 3\}$ is a S_3-set, where the left action is defined by $\sigma a = \sigma(a)$ for all $\sigma \in S_3$, $a \in I_3$. Find all distinct orbits of S_3. Find G_1, G_2, and G_3.
2. Let H be a subgroup of order 11 and index 4 of a group G. Prove that H is a normal subgroup of G.
3. Let H be a subgroup of a group G of index n. If H does not contain any nontrivial normal subgroups of G, prove that H is isomorphic to a subgroup of S_n.
4. Let $G = GL(2, \mathbb{R})$ and $S = \mathbb{R}^2$. Show that S is a G-set under the left action defined by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} (x, y) = (ax + by,\ cx + dy)$$

for all $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in G$, $(x, y) \in \mathbb{R}^2$.

5. Let G be a group of order 77 acting on a set S of 20 elements. Show that G must have a fixed point.
6. Let G be a group. The left action of G on the set G is defined by conjugation, i.e., $(g, x) \to gxg^{-1}$ for all $g, x \in G$. Show that the kernel of the homomorphism $\psi : G \to A(G)$ induced by this action is $Z(G)$.
7. Let G be a group of order 80 such that G has a subgroup of order 16. Show that G is not a simple group.
8. Show that a group of order 22 is not a simple group.
9. Show that there are no simple groups of orders 6, 10, 14, 26, 34, and 58.
10. Show that a group of order 8 cannot be a simple group.
11. Show that a simple group of order 63 cannot contain a subgroup of order 21.
12. Let G be a group of order 70 such that G has a subgroup of order 14. Show that G has a nontrivial normal subgroup.

Arthur Cayley(1821–1895) was born on August 16, 1821, in Cambridge, England. He was the second son. He entered Trinity College at the age of 17, as a pensioner. In 1842, he graduated as senior wrangler. Later he went to a law school and in 1849 he became a lawyer. As a lawyer, he made a comfortable living and in fourteen years, during which he practiced his law profession, he wrote approximately 300 mathematical papers.

In 1863, Cayley was elected to the new Sadlerian chair of pure mathematics at Cambridge, where he remained until his death. He died on January 26, 1895.

For most of his life, Cayley worked on mathematics, theoretical dynamics, and mathematical astronomy. In 1876, he published his only book, *Treatise* on *Elliptic* Functions. Cayley wrote 966 papers; there are thirteen volumes of his collected papers.

Cayley's mathematical style was terse. He usually wrote out his results and published them without delay. He, along with J. J. Sylvester, his lifelong friend, is considered to be the founder of invariant theory. He is also responsible for matrix theory. The square notation used for determinants is due to Cayley. He proved many important theorems of matrix theory, such as the Cayley-Hamilton theorem. He is one of the first mathematicians to consider geometry of more than three dimensions.

In 1854, Cayley published, "On the theory of groups depending on the symbolic equation $\theta^n = 1$." In this paper, he considered a group as a set of symbols, $1, \alpha, \beta, \ldots$, all of them different and such that the product of any two of them (no matter in what order), or the product of any one of them into itself, belongs to the set. This formulation of a group as a set of symbols and multiplications is different from the formulation considered by the earlier mathematicians. The paper is generally regarded as the earliest work on abstract group theory and Cayley is regarded as the founder of abstract group theory. He is best known for the theorem that every finite group is isomorphic to a suitable permutation group. In his article of 1854, he introduced a procedure for defining a finite group by listing its elements in the form of a multiplication table, known as a Cayley table. Cayley also proved a number of important theorems.

Chapter 6

Direct Product of Groups

6.1 External and Internal Direct Product

In Section 2.1, Exercise 25, we defined the direct product $G \times H$ of two groups G and H. In this section, we extend this concept to any finite family of groups and obtain their basic properties.

The notion of a direct product is used to factor a group into a product of smaller groups. This factorization gives structural properties of a group. In some cases, it allows for the complete characterization of a certain type of group. In Chapter 9, the concept of direct product is used to give a complete system of invariants for a finitely generated Abelian group, i.e., a finite set of positive integers which implies the isomorphism of any two finitely generated Abelian groups that have this set of integers.

Recall that $I_n = \{1, 2, \ldots, n\}$.

Let $\{G_i \mid i \in I_n\}$ be a family of groups. Let

$$G = G_1 \times G_2 \times \cdots \times G_n = \{(a_1, a_2, \ldots, a_n) \mid a_i \in G_i, i \in I_n\}.$$

Define $*$ on G as follows: for all $(a_1, a_2, \ldots, a_n)$, $(b_1, b_2, \ldots, b_n) \in G$

$$(a_1, a_2, \ldots, a_n) * (b_1, b_2, \ldots, b_n) = (a_1 b_1, a_2 b_2, \ldots, a_n b_n).$$

In the following theorem, we show that $*$ is a binary operation on G and that the set G together with the binary operation $*$ is a group. We also obtain several important properties of G.

Theorem 6.1.1 *Let $\{G_i \mid i \in I_n\}$ be a family of groups and $G = G_1 \times G_2 \times \cdots \times G_n$. Let e_i be the identity of G_i for all $i \in I_n$. Then $(G, *)$, where $*$ is defined above, is a group with $e = (e_1, e_2, \ldots, e_n)$ the identity element, and for all $(a_1, a_2, \ldots, a_n) \in G$,*

$$(a_1, a_2, \ldots, a_n)^{-1} = (a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1}).$$

Furthermore, let

$$H_i = \{(e_1, e_2, \ldots, e_{i-1}, a_i, e_{i+1}, \ldots, e_n) \mid a_i \in G_i\}$$

for all $i \in I_n$. Then the following assertions hold.

(i) H_i is a normal subgroup of G for all $i \in I_n$.

(ii) For all $a \in G$, a can be uniquely expressed as $a = h_1 h_2 \cdots h_n$, where $h_i \in H_i$, $i \in I_n$.

(iii) $H_i \cap (H_1 H_2 \cdots H_{i-1} H_{i+1} \cdots H_n) = \{e\}$ for all $i \in I_n$.

(iv) $G = H_1 H_2 \cdots H_n$.

Proof. First we note that $*$ is single-valued and if $(a_1, \ldots, a_n)$, $(b_1, \ldots, b_n) \in G$, then $(a_1, \ldots, a_n) * (b_1, \ldots, b_n) = (a_1 b_1, \ldots, a_n b_n) \in G$ since $a_i b_i \in G_i$ for all i. Thus, $*$ is a binary operation on G. We ask the reader to verify that $*$ is associative. Now $e = (e_1, e_2, \ldots, e_n) \in G$ and for all $a = (a_1, a_2, \ldots, a_n) \in G$,

$$\begin{aligned} ae &= (a_1, a_2, \ldots, a_n)(e_1, e_2, \ldots, e_n) \\ &= (a_1 e_1, a_2 e_2, \ldots, a_n e_n) \\ &= (a_1, a_2, \ldots, a_n) \\ &= a. \end{aligned}$$

Similarly, $ea = a$. Hence, e is the identity of G. To show that every element of G has an inverse in G, let $(a_1, a_2, \ldots, a_n) \in G$. Then $(a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1}) \in G$ since $a_i^{-1} \in G_i$ for all i and

$$\begin{aligned}(a_1, a_2, \ldots, a_n)(a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1}) &= (a_1a_1^{-1}, a_2a_2^{-1}, \ldots, a_na_n^{-1}) \\ &= (e_1, e_2, \ldots, e_n) \\ &= e.\end{aligned}$$

Similarly, $(a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1})(a_1, a_2, \ldots, a_n) = e$. Thus, every element of G has an inverse. Consequently, $(G, *)$ is a group. We also note that by the uniqueness of the inverse of an element

$$(a_1, a_2, \ldots, a_n)^{-1} = (a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1}).$$

(i) Let $i \in I_n$. Since $(e_1, e_2, \ldots, e_n) \in H_i$, $H_i \neq \emptyset$. Let $a = (e_1, \ldots, a_i, \ldots, e_n)$, $b = (e_1, \ldots, b_i, \ldots, e_n) \in H_i$. Then

$$\begin{aligned}ab^{-1} &= (e_1, \ldots, a_i, \ldots, e_n)(e_1, \ldots, b_i, \ldots, e_n)^{-1} \\ &= (e_1, \ldots, a_i, \ldots, e_n)(e_1, \ldots, b_i^{-1}, \ldots, e_n) \\ &= (e_1, \ldots, a_ib_i^{-1}, \ldots, e_n) \in H_i.\end{aligned}$$

Thus, H_i is a subgroup of G by Theorem 4.1.6. Let $g = (g_1, g_2, \ldots, g_n) \in G$. Then

$$\begin{aligned}gag^{-1} &= (g_1, g_2, \ldots, g_n)(e_1, \ldots, a_i, \ldots, e_n)(g_1, g_2, \ldots, g_n)^{-1} \\ &= (g_1, g_2, \ldots, g_ia_i, \ldots, g_n)(g_1^{-1}, g_2^{-1}, \ldots, g_n^{-1}) \\ &= (e_1, \ldots, g_ia_ig_i^{-1}, \ldots, e_n) \in H_i \text{ since } g_ia_ig_i^{-1} \in G_i.\end{aligned}$$

Hence, H_i is a normal subgroup of G.

(ii) Let $a = (a_1, a_2, \ldots, a_n) \in G$. Let $h_i = (e_1, \ldots, a_i, \ldots, e_n)$ for all $i \in I_n$. Then $a = h_1h_2 \cdots h_n$. To show that the representation of a is unique, let $a = k_1k_2 \cdots k_n$ be another representation of a, where $k_i \in H_i$ for all $i \in I_n$. Let $k_i = (e_1, \ldots, b_i, \ldots, e_n) \in H_i$ for all $i \in I_n$. Then

$$(a_1, a_2, \ldots, a_n) = h_1h_2 \cdots h_n = a = k_1k_2 \cdots k_n = (b_1, b_2, \ldots, b_n).$$

This implies that $a_i = b_i$ for all $i \in I_n$ and so $h_i = k_i$ for all $i \in I_n$. Hence, the representation of a is unique.

(iii) Suppose $a \in H_i \cap (H_1 \cdots H_{i-1}H_{i+1} \cdots H_n)$. Then $a \in H_i$ and

$$a \in H_1 \cdots H_{i-1}H_{i+1} \cdots H_n.$$

Since $a \in H_i$, $a = (e_1, \ldots, a_i, \ldots, e_n) \in H_i$ for some $a_i \in G_i$ and since

$$a \in H_1 \cdots H_{i-1}H_{i+1} \cdots H_n,$$

we have $a = h_1h_2 \cdots h_{i-1}h_{i+1} \cdots h_n$, where $h_j = (e_1, \ldots, a_j, \ldots, e_n) \in H_j$ for some $a_j \in G_j$. Thus,

$$(e_1, \ldots, a_i, \ldots, e_n) = a = h_1 \cdots h_{i-1}h_{i+1} \cdots h_n = (a_1, \ldots, a_{i-1}, e_i, a_{i+1,} \ldots, a_n).$$

This implies that $a_i = e_i$ for all $i \in I_n$. Hence,

$$H_i \cap (H_1H_2 \cdots H_{i-1}H_{i+1} \cdots H_n) = \{e\}.$$

(iv) The desired result follows from (ii). ■■

Definition 6.1.2 *The group G of Theorem* 6.1.1 *is called the* ***external direct product*** *of the groups G_i, $i = 1, 2, \ldots, n$.*

Theorem 6.1.1 motivates the following definition.

Definition 6.1.3 *Let G be a group and $\{N_i \mid i \in I_n\}$ be a family of normal subgroups of G. Then G is called the* ***internal direct product of*** *$N_1, N_2, \ldots, N_n$ if if $G = N_1N_2 \cdots N_n$ and $N_i \cap (N_1 \cdots N_{i-1}N_{i+1} \cdots N_n) = \{e\}$ for all $i \in I_n$.*

Let $G = G_1 \times G_2 \times \cdots \times G_n$ be the external direct product of the groups G_i. Let H_i be defined as in Theorem 6.1.1. Then G is the internal direct product of $H_1, H_2, \ldots, H_n$ by Theorem 6.1.1.

Theorem 6.1.4 *Let G be a group and $\{N_i \mid i \in I_n\}$ be a family of normal subgroups of G. Then G is an internal direct product of $\{N_i \mid i \in I_n\}$ if and only if for all $a \in G$, a can be uniquely expressed as $a = a_1a_2 \cdots a_n$, where $a_i \in N_i$, $i \in I_n$.*

Proof. Let G be an internal direct product of $\{N_i \mid i \in I_n\}$. Then $G = N_1N_2\cdots N_n$ and $N_i \cap (N_1\cdots N_{i-1}N_{i+1}\cdots N_n) = \{e\}$ for all $i \in I_n$. Then $N_i \cap N_j = \{e\}$ for all $i \neq j$ and hence $uv = vu$ for all $u \in N_i$ and for all $v \in N_j$ by Exercise 13 (page 93). Let $a = a_1a_2\cdots a_n = b_1b_2\cdots b_n$ be two representations of a, where a_i, $b_i \in N_i$, $i \in I_n$. Then

$$\begin{aligned} e &= a^{-1}a \\ &= (a_1a_2\cdots a_n)^{-1}(b_1b_2\cdots b_n) \\ &= a_n^{-1}a_{n-1}^{-1}\cdots a_1^{-1}b_1b_2\cdots b_n \\ &= a_1^{-1}b_1a_2^{-1}b_2\cdots a_n^{-1}b_n \end{aligned}$$

since for all $i \neq j$ if $u \in N_i$ and $v \in N_j$, then $uv = vu$. This implies that

$$b_i^{-1}a_i = a_1^{-1}b_1\cdots a_{i-1}^{-1}b_{i-1}a_{i+1}^{-1}b_{i+1}\cdots a_n^{-1}b_n \in N_i \cap N_1N_2\cdots N_{i-1}N_{i+1}\cdots N_n$$

for all $i \in I_n$. Since $N_i \cap N_1N_2\cdots N_{i-1}N_{i+1}\cdots N_n = \{e\}$, we must have $b_i^{-1}a_i = e$ or $a_i = b_i$ for all $i \in I_n$. Thus, a can be written uniquely as $a_1a_2\cdots a_n$, where $a_i \in N_i$, $i \in I_n$.

Conversely assume that for all $a \in G$, a can be uniquely expressed as $a = a_1a_2\cdots a_n$, where $a_i \in N_i$, $i \in I_n$.This implies that $G = N_1N_2\cdots N_n$. We now show that $N_i \cap (N_1\cdots N_{i-1}N_{i+1}\cdots N_n) = \{e\}$ for all $i \in I_n$. Let $i \in I_n$ and $a \in N_i \cap (N_1\cdots N_{i-1}N_{i+1}\cdots N_n)$. Then $a \in N_i$ and $a \in N_1\cdots N_{i-1}N_{i+1}\cdots N_n$. This implies that we can write $a = a_1a_2\cdots a_{i-1}a_{i+1}\cdots a_n$ for some $a_j \in N_j$, $j \in I_n\backslash\{i\}$. Hence,

$$ee\cdots a\cdots e = a = a_1a_2\cdots a_{i-1}ea_{i+1}\cdots a_n$$

are two representations of a, where $a_j \in N_j$, $j \in I_n\backslash\{i\}$. Since the representation of a is unique, $a = e$. Hence, $N_i \cap (N_1\cdots N_{i-1}N_{i+1}\cdots N_n) = \{e\}$. ■

In the following theorem, we show that if a group G is an internal direct product of a family of normal subgroups $\{N_i \mid i \in I_n\}$, then G can be viewed as an external direct product of the groups N_i's.

Theorem 6.1.5 *Let G be an internal direct product of a family of normal subgroups $\{N_i \mid i \in I_n\}$. Then*

$$G \simeq N_1 \times N_2 \times \cdots \times N_n.$$

Proof. Let $a \in G$. Then a can be expressed uniquely as $a = a_1a_2\cdots a_n$, where $a_i \in N_i$, $i \in I_n$. Define

$$f : G \to N_1 \times N_2 \times \cdots \times N_n$$

by

$$f(a) = (a_1, a_2, \ldots, a_n)$$

for all $a \in G$. From the definition of f, it follows that f is well defined and onto $N_1 \times N_2 \times \cdots \times N_n$. And from the uniqueness of the representation of a, it follows that f is one-one. We now show that f is a homomorphism. Let $a = a_1a_2\cdots a_n$ and $b = b_1b_2\cdots b_n$ be two elements of G, where a_i, $b_i \in N_i, i \in I_n$. Now $N_i \cap N_j = \{e\}$ for all $i \neq j$ and so $uv = vu$ for all $u \in N_i$, $v \in N_j$. This implies that

$$ab = a_1a_2\cdots a_nb_1b_2\cdots b_n = a_1b_1a_2b_2\cdots a_nb_n.$$

Thus,

$$\begin{aligned} f(ab) &= (a_1b_1, a_2b_2, \ldots, a_nb_n) \\ &= (a_1, a_2, \ldots, a_n)(b_1, b_2, \ldots, b_n) \\ &= f(a)f(b) \end{aligned}$$

and so f is a homomorphism. Consequently, $G \simeq N_1 \times N_2 \times \cdots \times N_n$. ■

Considering Theorem 6.1.5, let us agree to write $G = N_1 \times N_2 \times \cdots \times N_n$ when G is an internal direct product of a family of normal subgroups $\{N_i \mid i \in I_n\}$.

Worked-Out Exercises

♢ **Exercise 1** Let G and G_1 be groups and $f : G \to G_1$ be a homomorphism. Let H be a normal subgroup of G. Suppose that $f|_H : H \to G_1$ is an isomorphism of H onto G_1. Prove that $G = H \times \text{Ker } f$. Give an example to show that this result need not be true if H is not a normal subgroup.

Solution: Let $a \in G$. Then $f(a) \in G_1 = f(H)$. Thus, there exists $h \in H$ such that $f(a) = f(h)$. Now $f(a) = f(h)$ implies that $f(h^{-1}a) = e_1$ and hence $h^{-1}a \in \text{Ker } f$. Therefore, there exists $b \in \text{Ker } f$ such that $b = h^{-1}a$ or $a = hb$. Hence, $G = H\text{Ker } f$. Suppose $a \in H \cap \text{Ker } f$. Then $a \in H$ and $f(a) = e_1 = f(e)$. Since $f|_H$ is one-one, $f(a) = f(e)$ implies that $a = e$. Therefore, $H \cap \text{Ker } f = \{e\}$. Thus, H and Ker f are normal subgroups of G such that $G = H\text{Ker } f$ and $H \cap \text{Ker } f = \{e\}$. Consequently, $G = H \times \text{Ker } f$.

This result need not be true if H is not a normal subgroup of G. For let $G = S_3$ and $G_1 = \langle g' \rangle$ be such that $\circ(g') = 2$, i.e., G_1 is a cyclic group of order 2. Let $H = \langle (1\ 2) \rangle$. Define $f : G \to G_1$ by $f(e) = e$, $f(x) = e$ if x is an element of order 3, and $f(x) = g'$ if x is an element of order 2. Then $f\mid_H : H \to G_1$ is an isomorphism of H onto G_1. Now Ker $f = \{e, (1\ 2\ 3), (1\ 3\ 2)\} = \langle (1\ 2\ 3) \rangle$. But $G \neq H \times$ Ker f (see Exercise 14, page 127.)

♢ **Exercise 2** Let G be a group and H and K be subgroups of G such that $G = H \times K$. Let N be a normal subgroup of G such that $N \cap H = \{e\}$ and $N \cap K = \{e\}$. Prove that N is commutative.

Solution: Since $G = H \times K$, H and K are normal subgroups of G. Now for all $n \in N, h \in H, k \in K, nh = hn$, and $nk = kn$ by Exercise 13 (page 93). Let $a, b \in N$. Then there exist $h \in H, k \in K$ such that $b = hk$. Now $ab = a(hk) = (ah)k = (ha)k = h(ak) = h(ka) = (hk)a = ba$. Hence, N is commutative.

♢ **Exercise 3** Let G be a group and A and B be subgroups of G. If

(i) $G = AB$,

(ii) $ab = ba$ for all $a \in A$, $b \in B$, and

(iii) $A \cap B = \{e\}$,

prove that G is an internal direct product of A and B.

Solution: Let us first show that A and B are normal subgroups of G. For this, let $a \in A$, $g \in G$. There exist $c \in A$ and $b \in B$ such that $g = cb$ by (i). Now $gag^{-1} = (cb)a(cb)^{-1} = cbab^{-1}c^{-1} = cabb^{-1}c^{-1} = cac^{-1} \in A$. Hence, A is a normal subgroup of G. Similarly, B is a normal subgroup of G. Let $g \in G$. Then $g = ab$ for some $a \in A$, $b \in B$. Suppose $g = a_1b_1$, where $a_1 \in A$, $b_1 \in B$. Then $ab = a_1b_1$, which implies that $a_1^{-1}a = b_1b^{-1} \in A \cap B = \{e\}$. Thus, $a = a_1$ and $b = b_1$. Therefore, we find that every element g of G can be expressed uniquely as $g = ab$, $a \in A$, $b \in B$. Consequently, G is an internal direct product of A, B.

♢ **Exercise 4** Let G be a cyclic group of order mn, where m, n are positive integers such that $\gcd(m, n) = 1$. Show that $G \simeq \mathbb{Z}_m \times \mathbb{Z}_n$.

Solution: Since m divides $|G|$ and G is cyclic, there exists a unique cyclic subgroup A of G of order m by Theorem 4.2.11. Similarly, there exists a unique cyclic subgroup B of G of order n. Now $|A \cap B|$ divides $|A| = m$ and $|A \cap B|$ divides $|B| = n$. Since $\gcd(m, n) = 1$, $|A \cap B| = 1$. Thus, by Theorem 4.3.14,

$$|AB| = \frac{|A|\,|B|}{|A \cap B|} = \frac{mn}{1} = mn = |G|\,.$$

Since $AB \subseteq G$, $|AB| = |G|$, and G is finite, we must have $G = AB$. Hence, $G = AB$, $A \cap B = \{e\}$, and A and B are normal subgroups of G. Thus, $G = A \times B \simeq \mathbb{Z}_m \times \mathbb{Z}_n$.

♢ **Exercise 5** Let A and B be two cyclic groups of order m and n, respectively. Show that $A \times B$ is a cyclic group if and only if $\gcd(m, n) = 1$.

Solution: Let $A = \langle a \rangle$ for some $a \in A$ and $B = \langle b \rangle$ for some $b \in B$. Suppose $\gcd(m, n) = 1$. Let $g = (a, b)$. Then $g^{mn} = (a, b)^{mn} = (a^{mn},\ b^{mn}) = (e_A, e_B)$, where e_A denotes the identity of A and e_B denotes the identity of B. Suppose $\circ(g) = t$. Then $(a, b)^t = (e_A, e_B)$. This implies that $a^t = e_A$ and $b^t = e_B$. Thus, $m \mid t$ and $n \mid t$. Since $\gcd(m, n) = 1$, $mn \mid t$. Hence, mn is the smallest positive integer such that $g^{mn} = e$. Thus, $\circ(g) = mn$. Now $|A \times B| = mn$ and $A \times B$ contains an element g of order mn. As a result, $A \times B$ is cyclic. Conversely, assume that $A \times B$ is cyclic and $\gcd(m, n) = d \neq 1$. Let $(a, b) \in A \times B$. Then $\circ(a) \mid m$ and $\circ(b) \mid n$. Now $\frac{mn}{d} = \frac{m}{d}n = m\frac{n}{d}$ is an integer and $\frac{mn}{d} < mn$. Also,

$$(a, b)^{\frac{mn}{d}} = (a^{m\frac{n}{d}}, b^{n\frac{m}{d}}) = (e_A, e_B).$$

Hence, $A \times B$ does not contain any element of order mn. This implies that $A \times B$ is not cyclic, a contradiction. Therefore, $\gcd(m, n) = 1$.

♢ **Exercise 6** Show that $|\mathrm{Aut}(\mathbb{Z}_2 \times \mathbb{Z}_2)| = 6$.

Solution: First note that $\mathbb{Z}_2 \times \mathbb{Z}_2$ has four elements, $e = ([0], [0])$, $a = ([1], [0])$, $b = ([0], [1])$, $c = ([1], [1])$, and $\circ(a) = \circ(b) = \circ(c) = 2$. Let $f \in \mathrm{Aut}(\mathbb{Z}_2 \times \mathbb{Z}_2)$. Then $\circ(f(x)) = \circ(x)$ for all $x \in \mathbb{Z}_2 \times \mathbb{Z}_2$. Hence, f maps $\{a, b, c\}$ onto $\{a, b, c\}$. Thus, f is a permutation of $\{a, b, c\}$. Since there are only six permutations of $\{a, b, c\}$, it follows that $|\mathrm{Aut}(\mathbb{Z}_2 \times \mathbb{Z}_2)| \leq 6$. Now $a + b = c$, $a + c = b$, $b + c = a$, and $a + a = e = b + b = c + c$. Thus, any permutation of $\{a, b, c\}$ gives rise to an automorphism of $\mathbb{Z}_2 \times \mathbb{Z}_2$. For example, let $\alpha : a \to b$, $b \to c$, $c \to a$, and $e \to e$. Now $\alpha(a+b) = \alpha(c) = a$ and $\alpha(a)+\alpha(b) = b+c = a$. Therefore, $\alpha(a+b) = \alpha(a)+\alpha(b)$. Similarly, $\alpha(a + c) = \alpha(a) + \alpha(c)$, $\alpha(b + c) = \alpha(b) + \alpha(c)$, $\alpha(a + a) = \alpha(a) + \alpha(a)$, $\alpha(b + b) = \alpha(b) + \alpha(b)$, and $\alpha(c + c) = \alpha(c) + \alpha(c)$. Hence, α is an automorphism. Thus, $|\mathrm{Aut}(\mathbb{Z}_2 \times \mathbb{Z}_2)| = 6$.

Exercises

1. Prove that the direct product of two groups A and B is commutative if and only if both groups A and B are commutative.
2. Let A, B, C, and D be four groups such that $A \simeq C$ and $B \simeq D$. Show that $A \times B \simeq C \times D$.
3. Let G be a group such that $G = H_1 \times H_2 \times \cdots \times H_n$, where H_i is a subgroup of G. Let K_i be a normal subgroup of G such that $K_i \subseteq H_i$, $1 \leq i \leq n$. Let $K = K_1 \times K_2 \times \cdots \times K_n$. Show that
$$\frac{G}{K} \simeq \frac{H_1}{K_1} \times \frac{H_2}{K_2} \times \cdots \times \frac{H_n}{K_n}.$$
4. Let G_i be a group, $1 \leq i \leq n$. Show that
$$Z(G_1 \times G_2 \times \cdots \times G_n) = Z(G_1) \times Z(G_2) \times \cdots \times Z(G_n).$$
5. Let G be a group and H and K be subgroups of G such that $G = H \times K$. Show that $G/K \simeq H$ and $G/H \simeq K$.
6. Let G be a finite cyclic group of order mn, where m and n are relatively prime. Let H and K be subgroups of G such that $|H| = m$ and $|K| = n$. Show that $G = H \times K$.
7. Prove that $\mathrm{Aut}(\mathbb{Z}_2 \times \mathbb{Z}_2) \simeq S_3$.
8. Let G be a group and H and K be normal subgroups of G such that $G = HK$. Let $H \cap K = N$. Show that
$$G/N \simeq H/N \times K/N.$$
9. Prove that a finite Abelian group G is the internal direct product of subgroups H and K if and only if (i) $H \cap K = \{e\}$ and (ii) $|G| = |H|\,|K|$.
10. Show that the Klein 4-group is isomorphic to the direct product of a cyclic group of order 2 with itself.
11. Show that a cyclic group of order 4 cannot be expressed as an internal direct product of two subgroups of order 2.
12. Show that a cyclic group of order 8 cannot be expressed as an internal direct product of two subgroups of order 4 and 2, respectively.
13. Can the cyclic group $\mathbb{Z}_{12}$ be expressed as an internal direct product of two proper subgroups?
14. Show that S_3 cannot be written as a direct product of proper subgroups.
15. Show that D_4 cannot be expressed as an internal direct product of two proper subgroups.
16. Consider the groups $\mathbb{Z}_2 \times S_3$, $\mathbb{Z}_2 \times \mathbb{Z}_6$, and $\mathbb{Z}_{12}$. Are any two of these groups isomorphic? Is any one noncommutative?
17. Show that the additive group $(\mathbb{Z},+)$ cannot be expressed as an internal direct product of two nontrivial subgroups.
18. Show that the additive group $(\mathbb{Q},+)$ cannot be expressed as an internal direct product of two nontrivial subgroups.

Heinrich Weber (1842–1913) was born on May 5, 1892, in Heidelberg, Germany. In 1860, he studied mathematics and physics at the University of Heidelberg. He received his Ph.D. in 1863. He was appointed as extraordinary professor at the University of Heidelberg in 1869 and also taught at Edgenössische Polytechnikum in Zurich, the University of Königsberg, the Technische Hochschule in Charlottenburg, and the universities of Marburg, Göttingen, and Strasbourg.

Weber was a friend of Richard Dedekind and they often collaborated. Together they edited the work of Riemann in 1876. Herman Minkowski and David Hilbert were among Weber's students.

Weber's main research interests were in analysis and its applications to mathematical physics and number theory. He was encouraged by von Neumann to investigate physical problems and by Richelot to study algebraic functions. Along the lines of Jacobi, he worked on the theory of differential equations. He proved Abel's theorem in its most general form. He also worked on physical problems concerning heat, static and current electricity, the motion of rigid bodies in liquids, and electrolytic displacement.

Weber's most profound and penetrating work is in algebra and number theory. He, jointly with Dedekind, did work of fundamental importance on algebraic functions.

In 1891, Weber gave the "modern" definition of an abstract finite group. One of his outstanding accomplishments was the proof of Kronecker's theorem, which states that absolute Abelian fields are cyclotomic.

Weber was an enthusiastic and inspiring teacher who took great interest in educational questions. He died on May 17, 1913.

Chapter 7

Introduction to Rings

In the previous chapters, we investigated mathematical systems with one binary operation. There are many mathematical systems, called rings, with two binary operations. The notion of a ring is an outgrowth of such mathematical systems as the integers, rational numbers, real numbers, and complex numbers.

Although David Hilbert coined the term "ring," it was E. Noether who, under the influence of Hilbert, set down the axioms for rings. In 1914, Fraenkel gave the first definition of a ring. However, it is no longer commonly used.

As we shall see, a ring is a particular combination of a group and a semigroup. Hence, our previous work will prove helpful in our examination of rings. However, it is not enough to examine a set with two independent binary operations. In order to obtain the full power of the axiomatic approach, we need a dependency between the two operations—in particular, the distributive laws.

7.1 Basic Properties

This section parallels Chapter 2. Furthermore, we introduce several notations and definitions which will be used throughout the text.

Example 7.1.1 *Consider $\mathbb{Z}$, the set integer, together with the binary operations $+$, usual addition of numbers, and $\cdot$, usual multiplication of numbers. By Example 2.1.9, $(\mathbb{Z}, +)$ is a commutative group. Also by Example 2.1.10, $(\mathbb{Z}, \cdot)$ is a semigroup. Moreover, the distributive laws hold in $\mathbb{Z}$. That is, for any integers $a, b, c \in \mathbb{Z}$,*

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c) \text{ and } (b + c) \cdot a = (b \cdot a) + (c \cdot a).$$

Example 7.1.2 *Consider the set $\mathbb{Z}_n$ and the binary operations $+_n$ and $\cdot_n$as defined in Examples 2.1.9 and 2.1.10, respectively, where n is positive integer. As shown in Example 2.1.9, $(\mathbb{Z}_n, +_n)$ is a commutative group. By Example 2.1.10, $(\mathbb{Z}_n, \cdot_n)$ is a semigroup. Moreover, for any $[a], [b], [c] \in \mathbb{Z}_n$, we have*

$$\begin{aligned} [a] \cdot_n ([b] +_n [c]) &= [a] \cdot_n [b + c] \\ &= [a(b + c)] \\ &= [ab + ac] \\ &= [ab] +_n [ac] \\ &= ([a] \cdot_n [b]) +_n ([a] \cdot_n [c]). \end{aligned}$$

Similarly,

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

That is, the distributive laws hold in $(\mathbb{Z}_n, +_n, \cdot_n)$.

In the previous two examples, we considered a mathematical system with two binary operations. In general, the two binary operations are denoted by $+$ (addition) and $\cdot$ (multiplication). Under the binary operation $+$ the mathematical system is a commutative group, under the binary operation $\cdot$ the mathematical system is a semigroup, and $\cdot$ distributes over $+$. There are many such mathematical systems and such mathematical systems are called rings. More specifically, a ring is a mathematical system $(R, +, \cdot)$ such that $(R, +)$ is a commutative group, $(R, \cdot)$ is a semigroup, and the distributive laws hold, i.e., for all a, b, $c \in R$,

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c),$$

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

We denote the identity of $(R,+)$ by the symbol 0. The additive inverse of an element $a \in R$ is denoted by $-a$.

We now give a complete definition of a ring.

Definition 7.1.3 *A **ring** is an ordered triple $(R,+,\cdot)$ such that R is a nonempty set and $+$ and $\cdot$ are two binary operations on R satisfying the following axioms.*

(R1) $(a+b)+c = a+(b+c)$ for all $a,b,c \in R$.

(R2) $a+b = b+a$ for all $a,b \in R$.

(R3) There exists an element 0 in R such that $a+0=a$ for all $a \in R$.

(R4) For all $a \in R$, there exists an element $-a \in R$ such that

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

(R5) $(a\cdot b)\cdot c = a\cdot(b\cdot c)$ for all $a,b,c \in R$.

(R6) $a\cdot(b+c) = (a\cdot b)+(a\cdot c)$ for all $a,b,c \in R$.

(R7) $(b+c)\cdot a = (b\cdot a)+(c\cdot a)$ for all $a,b,c \in R$.

We call 0, the **zero element** of the ring $(R,+,\cdot)$.

During the development of the theory of rings, we will use the following conventions.

1. Multiplication is assumed to be performed before addition.
2. We write ab for $a\cdot b$.
3. We write $a-b$ for $a+(-b)$.
4. We refer to a ring $(R,+,\cdot)$ as a ring R.

Accordingly, $ab+c$ stands for $(a\cdot b)+c$, $ab+ac$ stands for $(a\cdot b)+(a\cdot c)$, $ab-ac$ stands for $(a\cdot b)+(-(a\cdot c))$, where $a,b,c \in R$.

Example 7.1.4 *(i) As shown in Examples 7.1.1 and 7.1.2, repsectively, $(\mathbb{Z},+,\cdot)$ and $(\mathbb{Z}_n,+_n,\cdot_n)$ are rings.*

(ii) It can be shown that $(\mathbb{Q},+,\cdot)$, $(\mathbb{R},+,\cdot)$, and $(\mathbb{C},+,\cdot)$ are rings.

*(iii) Consider $\mathbb{E}$, the set of even integers. Because addition and multiplication of even integers is an integer, we can show that $(\mathbb{E},+,\cdot)$ is a ring, called the **ring of even integers**. We leave the details as an exercise. However, note the 0 is the additive identity of the ring $\mathbb{E}$.*

The ring $(\mathbb{Z},+,\cdot)$ of Example 7.1.4(i) is called the **ring of integers**. This ring plays an important role in the study of ring theory. One of the basic problems in ring theory is to determine rings, which satisfy the same type of properties as the ring of integers.

Remark 7.1.5 *The ring $(\mathbb{Z}_n,+_n,\cdot_n)$ Example 7.1.4(i) is called the **ring of integers mod** n.*

Definition 7.1.6 *A ring R is called **commutative** if $ab=ba$ for all $a,b \in R$. A ring R which is not commutative is called a **noncommutative** ring.*

From the above definition, it follows that a ring R is commutative if and only if the semigroup $(R,\cdot)$ is commutative.

Example 7.1.7 *Because multipliation of numbers is commutative, it follows that $(\mathbb{E},+,\cdot)$, $(\mathbb{Z},+,\cdot)$, $(\mathbb{Q},+,\cdot)$, $(\mathbb{R},+,\cdot)$, and $(\mathbb{C},+,\cdot)$ are commutative rings. Also for any $[a],[b] \in \mathbb{Z}_n$, $[a]\cdot_n[b] = [b]\cdot_n[a]$. Hence, $(\mathbb{Z}_n,+_n,\cdot_n)$ is also a commutative ring.*

Definition 7.1.8 *For a ring R, the set $C(R) = \{a \in R \mid ab=ba \text{ for all } b \in R\}$ is called the **center** of R.*

It follows that a ring R is commutative if and only if $R = C(R)$.

Example 7.1.9 *Let $M_2(\mathbb{Z})$ denote the set of all 2×2 matrices over the ring of integers. Let $+$ and $\cdot$ denote the usual matrix addition and multiplication, respectively. Then $+$ and $\cdot$ are binary operations on $M_2(\mathbb{Z})$. It is easy to show that $(M_2(\mathbb{Z}), +, \cdot)$ is a ring. Note that $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ is the additive identity and for $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in M_2(\mathbb{Z})$,*

$$-\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} -a & -b \\ -c & -d \end{bmatrix}.$$

Now $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} \in M_2(\mathbb{Z})$ and

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix} \neq \begin{bmatrix} 23 & 34 \\ 31 & 46 \end{bmatrix} = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

Therefore, $M_2(\mathbb{Z})$ is not a commutative ring.

In $\mathbb{Z}$, we have $1n = n = n1$ for all $n \in \mathbb{Z}$. Such an element 1 is called an identity of the ring $\mathbb{Z}$. We thus, have the following definition.

Definition 7.1.10 *Let R be a ring. An element $e \in R$ is called an **identity** element if $ea = a = ae$ for all $a \in R$.*

Note that an identity element of a ring R (if it exists) is an identity element of the semigroup $(R, \cdot)$. Therefore, a ring cannot contain more than one identity element (Theorem 1.5.11). The identity element of a ring (if it exists) is denoted by 1.

Definition 7.1.11 *A ring R is called a **ring with identity** if it has an identity.*

Example 7.1.12 *The ring $\mathbb{Z}$ of integers is a ring with identity. The integer 1 is the identity element of $\mathbb{Z}$.*

Example 7.1.13 *Let n be a positive integer. The commutative ring $(\mathbb{Z}_n, +_n, \cdot_n)$ is with identity. The identity element is* [1].

Remark 7.1.14 *Let n be a positive integer. Note that the set $\mathbb{Z}_n$ has n elements. Therefore, by Example 7.1.13, it follows that for every positive integer n, there exists a commutative ring R with 1 such that the number of elements in R is n.*

Example 7.1.15 *Consider $\mathbb{E}$, the ring of even integers. In $\mathbb{E}$, there does not exist any element e such that $ex = x = xe$ for all $x \in \mathbb{E}$. Hence, $\mathbb{E}$, is a ring without identity.*

Example 7.1.16 *The ring $M_2(\mathbb{Z})$ of Example* 7.1.9 *is a ring with identity. The identity element of $M_2(\mathbb{Z})$ is $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.*

Example 7.1.17 *Let R denote the set of all functions $f : \mathbb{R} \to \mathbb{R}$. Define $+$, $\cdot$ on R by for all f, $g \in R$ and for all $a \in \mathbb{R}$,*

$$(f + g)(a) = f(a) + g(a),$$
$$(f \cdot g)(a) = f(a)g(a).$$

From the definition of $+$ and $\cdot$, it follows that $+$ and $\cdot$ are binary operations on R. Let $f, g, h \in R$. Then for all $a \in \mathbb{R}$, we have by using the associativity of $\mathbb{R}$ that $((f + g) + h)(a) = (f + g)(a) + h(a) = (f(a) + g(a)) + h(a) = f(a) + (g(a) + h(a)) = f(a) + (g + h)(a) = (f + (g + h))(a)$. Thus, $(f + g) + h = f + (g + h)$. This shows that $+$ is associative.

In a similar way, we can show that the other properties of a ring hold for R by using the fact that they hold for $\mathbb{R}$. Thus, $(R, +, \cdot)$ is a ring.

We note that the function $i_0 : \mathbb{R} \to \mathbb{R}$, where $i_0(a) = 0$ for all $a \in \mathbb{R}$, is the additive identity of R and the element $i_1 \in R$, where $i_1(a) = 1$ for all $a \in \mathbb{R}$, is the identity of R. Also, for all $f, g \in R$ and for all $a \in \mathbb{R}$,

$$(f \cdot g)(a) = f(a)g(a) = g(a)f(a) = (g \cdot f)(a).$$

Thus, for all $f, g \in R$, $f \cdot g = g \cdot f$. Consequently, $(R, +, \cdot)$ is a commutative ring with identity.

The addition and multiplication on R in Example 7.1.17 are the same as those encountered by the student in calculus.

Example 7.1.18 *Let $(G, *)$ be a commutative group and $Hom(G, G)$ be the set of all homomorphisms of G into itself. By Exercise 14, (page 104), the composition of two homomorphisms of G is again a homomorphism of G. Thus, $\circ$ is a binary operation on $Hom(G, G)$. Also, $\circ$ is associative by Theorem 1.4.13 and $i_G \in Hom(G, G)$ is the identity. Hence, $(Hom(G, G), \circ)$ is a semigroup with identity.*

We now define a suitable $+$ on $Hom(G, G)$ so that $(Hom(G, G), +, \circ)$ becomes a ring with identity. Define $+$ on $Hom(G, G)$ by for all $f, g \in Hom(G, G)$,

$$(f + g)(a) = f(a) * g(a) \text{ for all } a \in G.$$

(Note that $*$ is the binary operation of the group G.) Let $f, g \in Hom(G, G)$. From the definition of $+$, it follows that $f + g$ is a mapping from G into G. Let $a, b \in G$. Then

$$\begin{aligned} (f + g)(ab) &= f(ab) * g(ab) \\ &= (f(a) * f(b)) * (g(a) * g(b)) \\ &= f(a) * g(a) * f(b) * g(b) \\ &= (f + g)(a) * (f + g)(b). \end{aligned}$$

This shows that $f + g$ is a homomorphism from G into G. We omit the routine verification that $+$ is associative and commutative. Consider the mapping

$$f_e : G \to G$$

such that

$$f_e(a) = e$$

for all $a \in G$, where e is the identity of G. Note that f_e is a constant function, that maps each element of G to e. Now for all $a, b \in G$,

$$f_e(ab) = e = ee = f_e(a)f_e(b).$$

Thus, f_e is a homomorphism of G into G. Hence, $f_e \in Hom(G, G)$.

Let $g \in Hom(G, G)$ and $a \in G$. Now by the definition of f_e, we have $f_e(a) = e$. Thus,

$$\begin{aligned} (f_e + g)(a) &= f_e(a) * g(a) \\ &= e * g(a) \\ &= g(a). \end{aligned}$$

Also,

$$\begin{aligned} (g + f_e)(a) &= g(a) * f_e(a) \\ &= g(a) * e \\ &= g(a). \end{aligned}$$

Because a is an arbitrary element of G, we can now conclude that

$$f_e + g = g = g + f_e.$$

This implies that f_e is the identity of $(Hom(G, G), +)$.

We leave it as an exercise for the reader to verify that for any $f \in Hom(G, G)$, the mapping $-f$ defined by $(-f)(a) = f(a)^{-1}$ for all $a \in G$ is the additive inverse of f. Thus, $(Hom(G, G), +)$ is a commutative group.

We now show that the left distributive law holds. For any $a \in G$ and any elements $f, g, h \in Hom(G, G)$,

$$\begin{aligned} [f \circ (g + h)](a) &= f((g + h)(a)) \\ &= f(g(a) * h(a)) \\ &= f(g(a)) * f(h(a)) \\ &= (f \circ g)(a) * (f \circ h)(a) \\ &= (f \circ g + f \circ h)(a). \end{aligned}$$

Hence, $f \circ (g + h) = (f \circ g) + (f \circ h)$. The right distributive law holds similarly. Consequently, $(Hom(G, G), +, \circ)$ is a ring.

We now prove some elementary properties of rings.

Theorem 7.1.19 *Let R be a ring and $a, b, c \in R$. Then*
(i) $a0 = 0a = 0$,
(ii) $a(-b) = (-a)b = -(ab)$,
(iii) $(-a)(-b) = ab$,
(iv) $a(b-c) = ab - ac$ *and* $(b-c)a = ba - ca$.

Proof. (i) Observe that

$$a0 + a0 = a(0+0) = a0.$$

Thus,

$$\begin{array}{lll} & a0 + a0 = a(0+0) = a0 & \\ \Rightarrow & (a0 + a0) + (-(a0)) = a0 + (-(a0)) & \\ \Rightarrow & a0 + (a0 + (-(a0))) = 0 & \text{because } a0 + (-(a0)) = 0 \\ \Rightarrow & a0 + 0 = 0 & \text{because } a0 + (-(a0)) = 0 \\ \Rightarrow & a0 = 0 & \text{because } a0 + 0 = a0. \end{array}$$

Similarly, $0a = 0$.

(ii) We have

$$ab + a(-b) = a(b + (-b)) = a0 = 0.$$

Also

$$a(-b) + ab = a(-b + b) = a0 = 0.$$

Hence,

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

This implies that $a(-b)$ is an additive inverse of ab. Because the additive inverse of an element is unique, $a(-b) = -(ab)$. Similarly, $(-a)\ b = -(ab)$.

(iii) Using (ii), we have

$$(-a)(-b) = -(a(-b)) = -(-ab) = ab.$$

(iv) Because $b - c = b + (-c)$, $a(b-c) = a(b+(-c)) = ab + a(-c) = ab + (-(ac))$ (by (ii)) $= ab - ac$. Similarly, $(b-c)a = ba - ca$. ■

Corollary 7.1.20 *Let R be a ring with* 1. *Then $R \neq \{0\}$ if and only if the elements* 0 *and* 1 *are distinct.*

Proof. Suppose $R \neq \{0\}$. Let $a \in R$ be such that $a \neq 0$. Suppose $1 = 0$. Then $a = a1 = a0 = 0$, a contradiction. Thus, $1 \neq 0$. The converse follows because R has at least two distinct elements 0 and 1. ■

Convention From now on, we assume that the identity element 1 (if it exists) is different from the zero element of the ring.

From this convention, it follows that if R is a ring with 1, then R has at least two elements, namely the additive and multiplicatively identities.

Definition 7.1.21 *Let R be a ring with* 1. *An element $u \in R$ is called a* ***unit*** *(or an* ***invertible element****) if there exists $v \in R$ such that $uv = 1 = vu$.*

We note the following properties of invertible elements.

Theorem 7.1.22 *Let R be a ring with* 1 *and T be the set of all units of R. Then*
(i) $T \neq \emptyset$,
(ii) $0 \notin T$, *and*
(iii) $ab \in T$ *for all* $a, b \in T$.

Proof. (i) Because $1 \cdot 1 = 1 = 1 \cdot 1$, it follows that 1 is a unit. Thus, $1 \in T$. Hence, $T \neq \emptyset$.

(ii) Suppose that $0 \in T$. Then there exists $v \in R$ such that

$$0v = 1 = v0.$$

However, by Theorem 7.1.19(i), $0v = 0$. It now follows that $0 = 1$, which is a contradiction. Hence, $0 \notin T$.

(iii) Let $a, b \in T$. There exist $c, d \in R$ such that $ac = 1 = ca$ and $bd = 1 = db$. Now

$$(ab)(dc) = a(bd)c = a1c = ac = 1$$

and

$$(dc)(ab) = d(ca)b = d1b = db = 1.$$

Hence, $(ab)(dc) = 1 = (dc)(ab)$. Thus, ab is a unit, so $ab \in T$. ■

Definition 7.1.23 *(i) A ring R with 1 is called a **division ring (skew-field)** if every nonzero element of R is a unit.*

*(ii) A commutative division ring R is called a **field.***

Note that a ring R is a division ring (or skew-field) if and only if $(R\backslash\{0\}, \cdot)$ is a group. Therefore, if R is a division ring, then for all $a \in R$, $a \neq 0$, there exists a unique element, denoted by $a^{-1} \in R$, such that

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

We call a^{-1} the multiplicative inverse of a. In a similarly manner, a ring R is a field if and only if $(R\backslash\{0\}, \cdot)$ is a commutative group.

Example 7.1.24 *Consider $\mathbb{Z}$, the ring of integers. Let $a \in \mathbb{Z}$ be such that $a \neq 0$, $a \neq 1$, and $a \neq -1$. Now $a \cdot \frac{1}{a} = 1 = \frac{1}{a} \cdot a$. That is, the multiplicative inverse of a is $\frac{1}{a}$. However, $\frac{1}{a} \notin \mathbb{Z}$. (For example, the multiplicative inverse of 2 is $\frac{1}{2} \notin \mathbb{Z}$.). It follows that $\mathbb{Z}$ is not a field. Note that in $\mathbb{Z}$, the only invertible elements are 1 and -1.*

Example 7.1.25 *(i) From Example* 2.1.7, *$(\mathbb{Q}, +, \cdot)$ is a field, where $+$ and $\cdot$ are the usual addition and multiplication, respectively. $\mathbb{Q}$ is called the **field of rational numbers**.*

(ii) From Example 2.1.7, *$(\mathbb{R}, +, \cdot)$ is a field, where $+$ and $\cdot$ are the usual addition and multiplication, respectively. $\mathbb{R}$ is called the **field of real numbers**.*

(iii) From Example 2.1.7, *$(\mathbb{C}, +, \cdot)$ is a field, where $+$ and $\cdot$ are the usual addition and multiplication, respectively. $\mathbb{C}$ is called the **field of complex numbers**.*

The following example is due to William Rowan Hamilton. Due to physical considerations, Hamilton constructed a consistent algebra in which the commutative law of multiplication fails to hold. At the time, such a construction seemed inconceivable. His work and H.G. Grossman's work on hypercomplex number systems began the liberation of algebra. Their work encouraged other mathematicians to create algebras, which broke with tradition, e.g., algebras in which $ab = 0$ with $a \neq 0$, $b \neq 0$ and algebras with $a^n = 0$, where $a \neq 0$ and n is a positive integer.

Example 7.1.26 *Let $Q_{\mathbb{R}} = \{(a_1, a_2, a_3, a_4) \mid a_i \in \mathbb{R},\ i = 1, 2, 3, 4\}$. Define $+$ and $\cdot$ on $Q_{\mathbb{R}}$ as follows:*

$$(a_1, a_2, a_3, a_4) + (b_1, b_2, b_3, b_4) = (a_1 + b_1, a_2 + b_2, a_3 + b_3, a_4 + b_4)$$

$$\begin{aligned}(a_1, a_2, a_3, a_4) \cdot (b_1, b_2, b_3, b_4) \quad = \quad & (a_1b_1 - a_2b_2 - a_3b_3 - a_4b_4,\\ & a_1b_2 + a_2b_1 + a_3b_4 - a_4b_3,\\ & a_1b_3 + a_3b_1 + a_4b_2 - a_2b_4,\\ & a_1b_4 + a_2b_3 - a_3b_2 + a_4b_1).\end{aligned}$$

From the definition of $+$ and $\cdot$, it follows that $+$ and $\cdot$ are binary operations on $Q_{\mathbb{R}}$. Now $+$ is associative and commutative because addition is associative and commutative in $\mathbb{R}$. We also note that $(0, 0, 0, 0) \in Q_{\mathbb{R}}$ is the additive identity and if $(a_1, a_2, a_3, a_4) \in Q_{\mathbb{R}}$, then $(-a_1, -a_2, -a_3, -a_4) \in Q_{\mathbb{R}}$ and $-(a_1, a_2, a_3, a_4) = (-a_1, -a_2, -a_3, -a_4)$. Hence, $(Q_{\mathbb{R}}, +)$ is a commutative group. Similarly, $\cdot$ is associative and $(1, 0, 0, 0) \in Q_{\mathbb{R}}$ is the multiplicative identity.

*Let $(a_1, a_2, a_3, a_4) \in Q_{\mathbb{R}}$ be a nonzero element. Then $N = a_1^2 + a_2^2 + a_3^2 + a_4^2 \neq 0$ and $N \in \mathbb{R}$. Thus, $(a_1/N, -a_2/N, -a_3/N, -a_4/N) \in Q_{\mathbb{R}}$. We ask the reader to verify that $(a_1/N, -a_2/N, -a_3/N, -a_4/N)$ is the multiplicative inverse of (a_1, a_2, a_3, a_4). Thus, $Q_{\mathbb{R}}$ is a division ring and is called the ring of **real quaternions.** However, $Q_{\mathbb{R}}$ is not commutative because*

$$(0, 1, 0, 0) \cdot (0, 0, 1, 0) = (0, 0, 0, 1) \neq (0, 0, 0, -1) = (0, 0, 1, 0) \cdot (0, 1, 0, 0).$$

Therefore, $Q_{\mathbb{R}}$ is not a field.

Consider the ring $(\mathbb{Z}_8, +_8, \cdot_8)$. Now $[2], [4] \in \mathbb{Z}_8$ and $[2] \neq [0]$, $[4] \neq [0]$, and

$$[2] \cdot_8 [4] = [2 \cdot 4] = [8] = [0].$$

That is, $[2]$ and $[4]$ are nonzero, but their product is zero. There are other rings with such a property. This motivates the following definition.

Definition 7.1.27 *A nonzero element a in a ring R is called a **zero divisor** if there exists $b \in R$ such that $b \neq 0$ and either $ab = 0$ or $ba = 0$.*

Remark 7.1.28 *We do not call* 0 *a zero divisor.*

Remark 7.1.29 *An element cannot be a unit and zero divisor at the same time (Worked-Out Exercise 1, page 138). Thus, a field has no zero divisors.*

Consider $\mathbb{Z}$, the ring of integers. In this ring, if $a, b \in \mathbb{Z}$ and $a \neq 0$, $b \neq 0$, then $ab \neq 0$. From this it follows that $\mathbb{Z}$ has no zero divisors. However, as noted earlier $\mathbb{Z}$ is not a field. This motivates the following definition.

Definition 7.1.30 *Let R be a commutative ring with* 1. *Then R is called an* ***integral domain*** *if R has no zero divisors.*

Example 7.1.31 *The ring of integers $\mathbb{Z}$ is an integral domain.*

Example 7.1.32 *The ring $M_2(\mathbb{Z})$ is not an integral domain because it is noncommutative. Also, $M_2(\mathbb{Z})$ has zero divisors. For example, $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \in M_2(\mathbb{Z})$ and*

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

Example 7.1.33 *Let F be a field. Then F is a commutative ring with* 1. *By Worked-Out Exercise 1, page 138, if $a \in F$ and $a \neq 0$, then a is not a zero divisor. It now follows that F is an integral domain. Hence, every field is an integral domain.*

Example 7.1.34 *Consider $\mathbb{Z}[\sqrt{3}] = \{a+b\sqrt{3} \mid a, b \in \mathbb{Z}\}$. Then $\mathbb{Z}[\sqrt{3}]$ is an integral domain, where the operations $+$ and $\cdot$ are the usual operations of addition and multiplication. Note that $0 + 0\sqrt{3}$ is the additive identity of $\mathbb{Z}[\sqrt{3}]$ and $1 + 0\sqrt{3}$ is the multiplicative identity of $\mathbb{Z}[\sqrt{3}]$.*

Consider $\sqrt{3} \in \mathbb{Z}[\sqrt{3}]$. Suppose $\sqrt{3}$ is a unit in $\mathbb{Z}[\sqrt{3}]$. Then

$$(\sqrt{3})^{-1} = a + b\sqrt{3}$$

for some $a, b \in \mathbb{Z}$. If $a = 0$, then $(\sqrt{3})^{-1} = b\sqrt{3}$ or

$$1 = 3b,$$

which is a contradiction because this equation has no solutions in $\mathbb{Z}$. Therefore, $a \neq 0$. So $(\sqrt{3})^{-1} = a + b\sqrt{3}$ implies that

$$1 = a\sqrt{3} + 3b$$

or

$$\sqrt{3} = \frac{1 - 3b}{a} \in \mathbb{Q},$$

a contradiction. Hence, $\sqrt{3}$ is not a unit. We can now conclude that $\mathbb{Z}[\sqrt{3}]$ is not a field.

By arguments similar to the ones used in Example 7.1.34, we can show that the following sets are integral domains under the usual addition and multiplication.

$$\begin{aligned}
\mathbb{Z}[\sqrt{n}] &= \{a + b\sqrt{n} \mid a, b \in \mathbb{Z}\} \\
\mathbb{Z}[i\sqrt{n}] &= \{a + bi\sqrt{n} \mid a, b \in \mathbb{Z}\} \\
\mathbb{Z}[i] &= \{a + bi \mid a, b \in \mathbb{Z}\} \\
\mathbb{Q}[\sqrt{n}] &= \{a + b\sqrt{n} \mid a, b \in \mathbb{Q}\} \\
\mathbb{Q}[i\sqrt{n}] &= \{a + bi\sqrt{n} \mid a, b \in \mathbb{Q}\} \\
\mathbb{Q}[i] &= \{a + bi \mid a, b \in \mathbb{Q}\},
\end{aligned}$$

where n is a fixed positive integer and $i^2 = -1$. In fact, it can be shown that $\mathbb{Q}[\sqrt{n}], \mathbb{Q}[i\sqrt{n}]$, and $\mathbb{Q}[i]$ are fields.

Example 7.1.35 *The ring of even integers $\mathbb{E}$ is a commutative ring, without identity, and without zero divisors. Thus, $\mathbb{E}$ is not an integral domain.*

The ring appearing in the following example is sometimes useful in the construction of counterexamples.

Example 7.1.36 *Let $(R,+)$ be a commutative group. Define multiplication on R by $ab = 0$ for all $a, b \in R$, where 0 denotes the identity element of the group $(R,+)$. Then $(R,+,\cdot)$ is a ring called the **zero** ring. If R contains more than one element, then R is a commutative ring without 1 and every nonzero element of R is a zero divisor.*

The following theorem establishes a relation between zero divisors and the cancellation property of a ring.

Theorem 7.1.37 *Let R be a ring. If R has no zero divisors, then the cancellation laws hold, i.e., for all $a, b, c \in R$, $a \neq 0$, $ab = ac$ implies $b = c$ (**left cancellation law**) and $ba = ca$ implies $b = c$ (**right cancellation law**). If either cancellation law holds, then R has no zero divisors.*

Proof. Suppose R has no zero divisors. Let a, b, $c \in R$ be such that $ab = ac$ and $a \neq 0$. Then $ab - ac = 0$ or $a(b - c) = 0$. Becuase R has no zero divisors and $a \neq 0$, $a(b - c) = 0$ implies that $b - c = 0$ or $b = c$. Hence, the left cancellation law holds. Similarly, the right cancellation law holds.

Conversely, suppose one of the cancellation laws hold, say, the left, i.e., if a, b, $c \in R$, $a \neq 0$, then $ab = ac$ implies $b = c$.

Let a be a nonzero element of R and $b \in R$. Suppose $ab = 0$. Then $ab = a0$, from which $b = 0$ by canceling a.

Suppose $ba = 0$ and $b \neq 0$. Then $ba = b0$ and by canceling b, we obtain $a = 0$, a contradiction. Therefore, $b = 0$. Hence, R has no zero divisors.

Similarly, the right cancellation law implies that R has no zero divisors. ■

Definition 7.1.38 *A ring R is called a **finite** ring if R has only a finite number of elements; otherwise R is called an **infinite ring.***

The rings $\mathbb{Z}$ and $M_2(\mathbb{Z})$ are infinite.

Example 7.1.39 *Consider the ring $(\mathbb{Z}_n, +_n, \cdot_n)$. From Example 2.1.10, not every nonzero element of $\mathbb{Z}_n$ has an inverse. For example, suppose n is not prime, say, $n = 6$. Then $[4]$ has no multiplicative inverse in $\mathbb{Z}_6$. Also, $\mathbb{Z}_6$ has zero divisors. We have $[3] \neq [0] \neq [2]$. Because $[3] \cdot_6 [2] = [6] = [0]$, it follows that $[3]$ and $[2]$ are zero divisors. Thus, $\mathbb{Z}_6$ is not an integral domain and thus not a field. We can also conclude that $[2]$ and $[3]$ do not have multiplicative inverses because they are zero divisors.*

In the following result, we assume that the ring R is commutative. This assumption can be removed and the conclusion that R is a field remains valid. However, we have not developed the appropriate results to remove this assumption. We will prove the theorem in its most general form in Chapter 24.

Theorem 7.1.40 *A finite commutative ring R with more than one element and without zero divisors is a field.*

Proof. We must show that R has an identity and that every nonzero element of R is a unit.

Suppose that R has n elements. Let $a_1, a_2, \ldots, a_n$ be the distinct elements of R. Let $a \in R$, $a \neq 0$. Now $aa_i \in R$ for all i, so

$$\{aa_1, aa_2, \ldots, aa_n\} \subseteq R.$$

If $aa_i = aa_j$, then by Theorem 7.1.37, $a_i = a_j$. Therefore, the elements aa_1, $aa_2, \ldots,$ aa_n must be distinct. Because R has only n elements, it follows that

$$R = \{aa_1, aa_2, \ldots, aa_n\}.$$

This implies that $a \in \{aa_1, aa_2, \ldots, aa_n\}$. So one of the products must be equal to a, say, $aa_i = a$. Because R is commutative, we also have $a_i a = aa_i = a$.

We show that a_i is the identity of R. Let b be any element of R. Then $b \in \{aa_1, aa_2, \ldots, aa_n\}$. So there exists $a_j \in R$ such that $b = aa_j$. Thus,

$$\begin{aligned} ba_i &= a_i b && \text{(because } R \text{ is commutative)} \\ &= a_i(aa_j) && \text{(substituting for } b\text{)} \\ &= (a_i a)a_j \\ &= aa_j \\ &= b. \end{aligned}$$

This implies that a_i is the identity of R. We denote the identity of R by 1. Now $1 \in R = \{aa_1, aa_2, \ldots, aa_n\}$, so one of the products, say, aa_k, must equal 1, i.e., $aa_k = 1$. By commutativity, $a_k a = aa_k = 1$. Hence, every nonzero element is a unit. Consequently, R is a field. ■

The following corollary is immediate from Theorem 7.1.40.

Corollary 7.1.41 *Every finite integral domain is a field.* ■

In Example 2.1.10, we showed that a nonzero element $[a]$ of $\mathbb{Z}_n$ has an inverse if and only if $\gcd(a, n) = 1$. Thus, the following corollary is an immediate consequence of this fact. We leave the details as an exercise.

Corollary 7.1.42 *Let n be a positive integer. Then $\mathbb{Z}_n$ is a field if and only if n is prime.* ■

Let R be a ring and $a \in R$. Then for any integer n, define na as follows:

$$\begin{array}{rcll} 0a & = & 0 & \\ na & = & a + (n-1)a & \text{if } n > 0 \\ na & = & (-n)(-a) & \text{if } n < 0. \end{array}$$

We emphasize that na is not a multiplication of elements of R because R may not contain $\mathbb{Z}$. We have the following properties holding for any $a, b \in R$ and any $m, n \in \mathbb{Z}$:

$$\begin{array}{rcl} (m+n)a & = & ma + na, \\ m(a+b) & = & ma + mb, \\ (mn)a & = & m(na), \\ m(ab) & = & (ma)b = a(mb), \\ (ma)(nb) & = & mn(ab). \end{array}$$

The proofs of the above properties can be obtained by induction and the defining conditions of a ring.

We close this chapter by introducing the concept of the characteristic of a ring and proving its basic properties.

Definition 7.1.43 *Let R be a ring. If there exists a positive integer n such that for all $a \in R$, $na = 0$, then the smallest such positive integer is called the characteristic of R. If no such positive integer exists, then R is said to be of* ***characteristic zero****.*

Example 7.1.44 *The ring $\mathbb{Z}_n$, $n = 1, 2, 3, \ldots$, has characteristic n. Note that in $\mathbb{Z}_6$, $3[2] = [6] = [0]$ and $2[3] = [6] = [0]$. However, 6 is the smallest positive integer such that $6[a] = [0]$ for all $[a] \in \mathbb{Z}_6$. In particular, $[1]$ has additive order 6.*

Example 7.1.45 *The rings $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ have characteristic 0.*

Example 7.1.46 *Let X be a nonempty set and $\mathcal{P}(X)$ the power set of X. Then $(\mathcal{P}(X), \Delta, \cap)$ is a commutative ring with 1, where Δ is the operation "symmetric difference." In this example, Δ acts as $+$ and $\cap$ acts as $\cdot$. Now for all $A \in \mathcal{P}(X)$, $2A = A\Delta A = (A \backslash A) \cup (A \backslash A) = \emptyset$. Thus, $\mathcal{P}(X)$ has characteristic 2.*

Theorem 7.1.47 *Let R be a ring with 1. Then R has characteristic $n > 0$ if and only if n is the least positive integer such that $n1 = 0$.*

Proof. Suppose that R has characteristic $n > 0$. Then $na = 0$ for all $a \in R$, so in particular, $n1 = 0$.

Suppose that $m1 = 0$ for some m such that $0 < m < n$. Then

$$ma = m(1a) = (m1)a = 0a = 0$$

for all $a \in R$. However, this contradicts the minimality of n. Hence, n is the smallest positive integer such that $n1 = 0$.

Conversely, suppose n is the smallest positive integer such that $n1 = 0$. Then for all $a \in R$,

$$na = n(1a) = (n1)a = 0a = 0.$$

By the minimality of n for 1, n must be the characteristic of R. ■

Theorem 7.1.48 *The characteristic of an integral domain R is either zero or a prime.*

Proof. If there does not exist a positive integer n such that $na = 0$ for all $a \in R$, then R is of characteristic zero.

Suppose there exists a positive integer n such that $na = 0$ for all $a \in R$. Let m be the smallest positive integer such that $ma = 0$ for all $a \in R$. That is, the characteristic of R is m. Then

$$m1 = 0.$$

Suppose m is not prime. Then there exist integers m_1, m_2 such that $1 < m_1 < m, 1 < m_2 < m$, and $m = m_1m_2$. Hence,

$$0 = (m_1m_2)1 = (m_11)(m_21).$$

Because R has no zero divisors, either $m_11 = 0$ or $m_21 = 0$. This contradicts the minimality of m. Hence, m is prime. ■

Worked-Out Exercises

♢ **Exercise 1** Let R be a ring.

(a) Let R be with 1. Let $a \in R$ be such that a has an inverse. Show that a cannot be a zero divisor.

An element $a \in R$ is called **idempotent** if $a^2 = a$ and **nilpotent** if $a^n = 0$ for some positive integer n.

(b) Let $a \in R$ be a nonzero idempotent. Show that a is not nilpotent.

(c) Let R be with 1 and suppose R has no zero divisors. Show that the only idempotents in R are 0 and 1.

Solution: (a) There exists $b \in R$ such that $ab = 1 = ba$. Suppose that a is a zero divisor. Then there exists $c \in R$, $c \neq 0$, such that $ac = 0$. Thus, $0 = b0 = b(ac) = (ba)c = c$, which is a contradiction. Hence, a is not a zero divisor.

(b) From the hypothesis, $a^2 = a$. By induction, $a^n = a$ for all positive integers n. Suppose a is nilpotent. Then $a^m = 0$ for some positive integer m, so $a = a^m = 0$, which is a contradiction, so a is not nilpotent.

(c) Clearly 0 and 1 are idempotent elements. Let $e \in R$ be an idempotent. Then $e^2 = e$, so $e(e-1) = 0$. Because R has no zero divisors, either $e = 0$ or $e - 1 = 0$, i.e., either $e = 0$ or $e = 1$. Therefore, the only idempotents of R are 0 and 1.

♢ **Exercise 2** Determine positive integers n such that $\mathbb{Z}_n$ has no nonzero nilpotent elements.

Solution: We claim that n is a square free integer, i.e., $n = p_1p_2 \cdots p_k$, where the p_i's are distinct primes.

Suppose that $n = p_1p_2 \cdots p_k$, p_i's are distinct primes. Let $[a] \in \mathbb{Z}_n$ be nilpotent. Then $[a]^m = [0]$ for some integer m. Hence, n divides a^m, so $p_1p_2 \cdots p_k$ divides a^m. Then $p_i \mid a^m$ for all $i = 1, 2, \ldots, k$. Because the p_i's are prime, $p_i \mid a$ for all $i = 1, 2, \ldots, k$. Because $p_1, p_2, \ldots, p_k$ are distinct primes, we must have $p_1p_2 \cdots p_k \mid a$, i.e., $n \mid a$, so $[a] = [0]$. This implies that $\mathbb{Z}_n$ has no nonzero nilpotent elements. Conversely, suppose that $\mathbb{Z}_n$ has no nonzero nilpotent elements. Let $n = p_1^{m_1}p_2^{m_2} \cdots p_k^{m_k}$, where the p_i's are distinct primes and $m_i \geq 1$. Let $m = \max\{m_1, m_2, \ldots, m_k\}$. Now $[p_1p_2 \cdots p_k]^m = [p_1{}^m p_2^m \cdots p_k^m] = [0]$ because $n \mid (p_1^m p_2^m \cdots p_k^m)$. Also, because $\mathbb{Z}_n$ has no nonzero nilpotent elements, $[p_1p_2 \cdots p_k] = [0]$. Hence, $n \mid (p_1 \cdots p_k)$, so $(p_1^{m_1}p_2^{m_2} \cdots p_k^{m_k}) \mid (p_1 \cdots p_k)$. Thus, $m_i \leq 1$ for all $i = 1, 2, \ldots, k$. Hence, $m_i = 1$ for all $i = 1, 2, \ldots, k$, so n is a square free integer.

♢ **Exercise 3** Show that the number of idempotent elements in $\mathbb{Z}_{mn}$, where $m > 1$, $n > 1$, and m and n are relatively prime, is at least 4.

Solution: Clearly, [0] and [1] are idempotent elements. Because m and n are relatively prime, there exist integers a and b such that $am + bn = 1$. We now show that n does not divide a and m does not divide b. Suppose that $n \mid a$. Then $a = nr$ for some integer r. Thus, $n(rm + b) = nrm + nb = am + nb = 1$. This implies that $n = 1$, which is a contradiction. Therefore, n does not divide a and similarly m does not divide b. Now $m^2a = m(1 - nb)$. This implies that $[m^2a] = [m]$. Hence, $[ma]^2 = [ma]$. If $[ma] = [0]$, then $mn \mid ma$, so $n \mid a$, which is a contradiction. Consequently, $[ma] \neq [0]$. If $[ma] = [1]$, then $mn \mid (ma - 1)$. Hence, $ma + mnt = 1$ for some integer t. Thus, $m(a + nt) = 1$. This implies $m = 1$, which is a contradiction. Hence, $[ma] \neq [1]$. Thus, $[ma]$ is an idempotent such that $[ma] \neq [0]$ and $[ma] \neq [1]$. Similarly, $[nb]$ is an idempotent such that $[nb] \neq [0]$ and $[nb] \neq [1]$. Clearly $[ma] \neq [nb]$. Thus, we find that [0], [1], $[ma]$, and $[nb]$ are idempotent elements of $\mathbb{Z}_{mn}$.

♢ **Exercise 4** Determine the positive integers n such that $\mathbb{Z}_n$ has no idempotent elements other than [0] and [1].

Solution: We show that $n = p^r$ for some prime p and some integer $r > 0$.

First assume that $n = p^r$ for some prime p and some positive integer r and $[x] \in \mathbb{Z}_n$ be an idempotent. Then $[x]^2 = [x]$. Thus, $p^r \mid (x^2 - x)$ or $p^r \mid x(x-1)$. Because x and $x-1$ are relatively prime, $p^r \mid x$ or $p^r \mid (x-1)$. If $p^r \mid x$, then $[x] = [0]$ and if $p^r \mid (x-1)$, then $[x] = [1]$. Thus, [0] and [1] are the only two idempotent elements. Conversely, suppose that [0] and [1] are the only two idempotent elements. Let $n = p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}$, where the p_i's are distinct primes, $m_i \geq 1$, and $k > 1$. Let $t = p_1^{m_1}$ and $s = p_2^{m_2} \cdots p_k^{m_k}$. Then t and s are relatively prime and $n = ts$. By Worked-Out Exercise 3, $\mathbb{Z}_n = \mathbb{Z}_{ts}$ must have at least four idempotents, which is a contradiction. Therefore, $k = 1$. Thus, $n = p^r$ for some prime p and some positive integer r.

Exercise 5 Let R be a ring. Show that the following conditions are equivalent.

(i) R has no nonzero nilpotent elements.

(ii) For all $a \in R$, if $a^2 = 0$, then $a = 0$.

Solution: (i)⇒(ii) Let $a \in R$ and $a^2 = 0$. If $a \neq 0$, then a is a nonzero nilpotent element of R, a contradiction. Thus, $a = 0$.

(ii)⇒(i) Let $a \in R$ be such that $a^n = 0$ for some positive integer n. Suppose $a \neq 0$. Let n be the smallest positive integer such that $a^n = 0$. Suppose n is even, say, $n = 2m$ for some positive integer m. Then $(a^m)^2 = a^{2m} = 0$, so $a^m = 0$, contradicting the minimality of n. Suppose n is odd. If $n = 1$, then $a = 0$, a contradiction. Therefore, $n > 1$. Suppose $n = 2m+1$. Then $m+1 < n$. Thus, $a^{2m+2} = a^{2m+1}a = a^n a = 0$. This implies that $a^{m+1} = 0$, which is a contradiction of the minimality of n. Hence, R has no nonzero nilpotent elements.

♢ **Exercise 6** An element e of a ring R is called a **left (right) identity**, if $ea = a$ ($ae = a$) for all $a \in R$. Show that if a ring R has a unique left identity e, then e is also the right identity of R and hence the identity of R.

Solution: Let e be the unique left identity of R. Then $ex = x$ for all $x \in R$. Let $x \in R$. Now $(xe - x + e)x = xex - xx + ex = xx - xx + x = x$. This implies that $xe - x + e$ is a left identity. Because e is the unique left identity, $xe - x + e = e$, so $xe = x$. Thus, e is a right identity.

Exercise 7 Let R be a commutative ring with 1 and $a, b \in R$. Suppose that a is invertible and b is nilpotent. Show that $a + b$ is invertible. Also, show that if R is not commutative, then the result may not be true.

Solution: There exists $c \in R$ such that $ac = 1 = ca$ and there exists a positive integer n such that $b^n = 0$. Let $d = c - c^2b + c^3b^2 + \cdots + (-1)^{n+1}c^n b^{n-1}$. Now $(a+b)d = ac - ac^2b + ac^3b^2 + \cdots + (-1)^{n+1}ac^n b^{n-1} + bc - bc^2b + bc^3b^2 + \cdots + (-1)^{n+1}bc^n b^{n-1} = 1 - cb + c^2b^2 + \cdots + (-1)^{n+1}c^{n-1}\ b^{n-1} + bc - c^2b^2 + c^3b^3 + \cdots + (-1)^{n+1}c^n b^n = 1$. Similarly, $d(a+b) = 1$. Hence, $a + b$ is invertible.

Consider the ring $M_2(\mathbb{Z})$. Let $a = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ and $b = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Then a is invertible and b is nilpotent. Now $a + b = \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}$. Clearly $a + b$ is a nonzero nilpotent element. Hence, $a + b$ is not invertible.

Exercises

1. In the rings $\mathbb{Z}_8$ and $\mathbb{Z}_6$, find the following elements:
   (i) the units, (ii) the nilpotent elements, and (iii) the zero divisors.
2. Let R be the set of all 2×2 matrices over the field of complex numbers of the form $\begin{bmatrix} z_1 & z_2 \\ -\overline{z}_2 & \overline{z}_1 \end{bmatrix}$, where $\overline{z}$ denotes the complex conjugate of the complex number z. Show that $(R, +, \cdot)$ is a division ring, where $+$ and $\cdot$ are the usual matrix addition and matrix multiplication, respectively. Is R a field?
3. Let R be a ring with 1. Prove that
   (i) $(-1)a = -a = a(-1)$ and $(-1)(-1) = 1$,
   (ii) if a is a unit in R, then $-a$ is a unit in R and $(-a)^{-1} = -(a^{-1})$.
4. Prove that a ring R is commutative if and only if $(a+b)^2 = a^2 + 2ab + b^2$ for all $a, b \in R$.
5. Prove that a ring R is commutative if and only if $a^2 - b^2 = (a+b)(a-b)$ for all $a, b \in R$.
6. Let R be a ring. If $a^3 = a$ for all $a \in R$, prove that R is commutative.
7. Let R be a commutative ring and $a, b \in R$. Prove that for all $n \in \mathbb{N}$,

$$(a+b)^n = a^n + \binom{n}{1}a^{n-1}b + \cdots + \binom{n}{r}a^{n-r}b^r + \cdots + \binom{n}{n-1}ab^{n-1} + b^n.$$

8. If a and b are elements of a ring and m and n are integers, prove that
   (i) $(na)(mb) = (nm)(ab)$,
   (ii) $n(ab) = (na)b = a(nb)$,
   (iii) $n(-a) = (-n)a$.
9. If R is an integral domain of prime characteristic p, prove that $(a+b)^p = a^p + b^p$ for all a, $b \in R$.
10. Let R be a ring with 1 and without zero divisors. Prove that for all $a, b \in R$, $ab = 1$ implies $ba = 1$.
11. Let R be a ring with 1. If a is a nilpotent element of R, prove that $1 - a$ and $1 + a$ are units.
12. Let R be a division ring and $a, b \in R$. Show that if $ab = 0$, then either $a = 0$ or $b = 0$.
13. Let $a \in R$ be an idempotent element. Show that $(1 - a)ba$ is nilpotent for all $b \in R$.
14. Find all idempotent elements of the ring $M_2(\mathbb{R})$.
15. Let R be a ring with 1. Let $0 \neq a \in R$. If there exist two distinct elements b and c in R such that $ab = ac = 1$, show that there are infinitely many elements x in R such that $ax = 1$. (*American Mathematical Monthly* 70(1961) 315).
16. Let R be an integral domain and $a, b \in R$. Let $m, n \in \mathbb{Z}$ be such that m and n are relatively prime. Prove that $a^m = b^m$ and $a^n = b^n$ imply that $a = b$.
17. Let R and R' be rings. Define $+$ and $\cdot$ on $R \times R'$ by for all $(a, b), (c, d) \in R \times R'$

$$(a, b) + (c, d) = (a + c, b + d) \text{ and } (a, b) \cdot (c, d) = (a \cdot c, b \cdot d).$$

   (i) Prove that $(R \times R', +, \cdot)$ is a ring. This ring is called the **direct sum** of R and R' and is denoted by $R \oplus R'$.
   (ii) If R and R' are commutative with identity, prove that $R \oplus R'$ is commutative with identity.
18. Extend the notion of direct sum in Exercise 17 to any finite number of rings.
19. Prove that the characteristic of a finite ring R divides $|R|$.
20. Let R be a ring with 1. Prove that the characteristic of the matrix ring $M_2(R)$ is the same as that of R.
21. If p is a prime integer, prove that $(p-1)! \equiv_p -1$.
22. In the following exercises, write the proof if the statement is true; otherwise, give a counterexample.
   (i) In a ring R, if a and b are idempotent elements, then $a + b$ is an idempotent element.
   (ii) In a ring R, if a and b are nilpotent elements, then $a + b$ is a nilpotent element.
   (iii) Every finite ring with 1 is an integral domain.
   (iv) There exists a field with seven elements.
   (v) The characteristic of an infinite ring is always 0.
   (vi) An element of a ring R which is idempotent, but not a zero divisor, is the identity element of R.
   (vii) If a and b are two zero divisors, then $a + b$ is also a zero divisor in a ring R.
   (viii) In a finite field F, $a^2 + b^2 = 0$ implies $a = 0$ and $b = 0$ for all $a, b \in F$.
   (ix) In a field F, $(a + b)^{-1} = a^{-1} + b^{-1}$ for all nonzero elements a, b such that $a + b \neq 0$.
   (x) There exists a field with six elements.

7.2 Some Important Rings

In this section, we introduce two important rings and study some of their basic properties.

Boolean Rings

We recall that in Worked-Out Exercise 1 (page 138), an element x of a ring R is called an **idempotent** element if $x^2 = x$. The zero element and identity element of a ring are idempotent elements. In the ring $\mathbb{Z}$, the only idempotent elements are 0 and 1. There exist rings, which contain idempotent elements different from 0 and 1. For example, in $M_2(\mathbb{Z})$, $\begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix}$ is an idempotent element.

Definition 7.2.1 *A ring R with 1 is called a **Boolean ring** if every element of R is an idempotent.*

Example 7.2.2 *(i) $\mathbb{Z}_2$ is a Boolean ring.*
(ii) The ring $\mathcal{P}(X)$ of Example 7.1.46 *is a Boolean ring because for all $A \in \mathcal{P}(X)$, $A \cap A = A$.*

Theorem 7.2.3 *Let R be a Boolean ring. Then the characteristic of R is 2 and R is commutative.*

Proof. First we show that R is of characteristic 2. Let $x \in R$. Now $x + x = (x + x)^2 = (x + x)(x + x) = x(x + x) + x(x + x) = x^2 + x^2 + x^2 + x^2 = x + x + x + x$. This implies that $2x = 4x$, so $0 = 2x$. Hence, $2 \cdot 1 = 0$ because x was arbitrary. It follows that the characteristic of R is 2 by Theorem 7.1.47. To show R is commutative, let $x, y \in R$. Then $x + y = (x + y)^2 = (x + y)(x + y) = x^2 + xy + yx + y^2 = x + xy + yx + y$. This implies that $0 = xy + yx$. Hence, $xy = xy + 0 = xy + xy + yx$ or $xy = 2xy + yx = yx$ because $2xy = 0$. Thus, R is commutative. ■

Regular Rings

An element x of a ring R is called a **regular element** if there exists $y \in R$ such that $x = xyx$.

Definition 7.2.4 *A ring R is called a **regular ring** if every element of R is regular.*

In the ring $\mathbb{Z}$, the only regular elements are $0, 1$, and -1. Thus, $\mathbb{Z}$ is not a regular ring.

Example 7.2.5 *Let R be a division ring and $x \in R$. If $x = 0$, then $x = xxx$. Suppose $x \neq 0$. Then $xx^{-1} = 1$, so $x = xx^{-1}x$. Thus, R is a regular ring.*

From the definition of a Boolean ring, it follows that every Boolean ring is a regular ring. The field $\mathbb{R}$ is a regular ring, but not a Boolean ring.

Example 7.2.6 *Consider $\mathbb{R}$, the field of real numbers and*

$$\mathbb{R}\times\mathbb{R}=\{(x, y) \mid x, y \in \mathbb{R}\}.$$

Define $+$ and $\cdot$ on $\mathbb{R}\times\mathbb{R}$ by

$$\begin{aligned}(x, y) + (z, w) &= (x + z, y + w)\\ (x, y) \cdot (z, w) &= (xz, yw)\end{aligned}$$

for all $x, y, z, w \in \mathbb{R}$. Then $\mathbb{R}\times\mathbb{R}$ is a commutative ring with identity. Now $(1, 0), (0, 1) \in \mathbb{R}\times\mathbb{R}$ and $(1, 0)(0, 1) = (0, 0)$. This shows that $\mathbb{R}\times\mathbb{R}$ contains zero divisors, so $\mathbb{R}\times\mathbb{R}$ is not a field. We claim that $\mathbb{R}\times\mathbb{R}$ is regular. Let $(x, y) \in \mathbb{R}\times\mathbb{R}$. If $x = 0 = y$, then $(x, y)(x, y)(x, y) = (x, y)$. If $x \neq 0$ and $y \neq 0$, then $(x, y)(x^{-1}, y^{-1})(x, y) = (x, y)$. If $x = 0$, but $y \neq 0$, then $(x, y)(x, y^{-1})(x, y) = (x, y)$. Similarly, if $x \neq 0$ and $y = 0$, then $(x, y)(x^{-1}, y)(x, y) = (x, y)$. Thus, in any case, (x, y) is a regular element. Hence, $\mathbb{R}\times\mathbb{R}$ is a regular ring.

Example 7.2.7 *Let $M_2(\mathbb{R})$ be the set of all 2×2 matrices over $\mathbb{R}$. Now $M_2(\mathbb{R})$ is a noncommutative ring with 1, where $+$ and $\cdot$ are the usual matrix addition and multiplication, respectively. We show that $M_2(\mathbb{R})$ is a regular ring. Let $A = \begin{bmatrix} x & y \\ z & w \end{bmatrix} \in M_2(\mathbb{R})$.*

***Case 1**: $xw - zy \neq 0$. Then $B = \begin{bmatrix} \frac{w}{xw-zy} & \frac{-y}{xw-zy} \\ \frac{-z}{xw-zy} & \frac{x}{xw-zy} \end{bmatrix} \in M_2(\mathbb{R})$ and $A = ABA$.*

***Case 2**: $xw - zy = 0$.*

***Subcase 2a**: x, y, z, w are all zero. In this case, $A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, so for any $B \in M_2(\mathbb{R})$, $ABA = A$.*

***Subcase 2b**: x, y, z, w are not all zero. Suppose $x \neq 0$ and let $B = \begin{bmatrix} \frac{1}{x} & 0 \\ 0 & 0 \end{bmatrix}$. Then*

$$\begin{aligned} ABA &= \begin{bmatrix} x & y \\ z & w \end{bmatrix}\begin{bmatrix} \frac{1}{x} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x & y \\ z & w \end{bmatrix}\\ &= \begin{bmatrix} 1 & 0 \\ \frac{z}{x} & 0 \end{bmatrix}\begin{bmatrix} x & y \\ z & w \end{bmatrix}\\ &= \begin{bmatrix} x & y \\ z & \frac{zy}{x} \end{bmatrix} = \begin{bmatrix} x & y \\ z & w \end{bmatrix}\end{aligned}$$

because $xw - zy = 0$ *and* $x \neq 0$ *implies* $w = \frac{zy}{x}$. *If* $y \neq 0$, *then let* $B = \begin{bmatrix} 0 & 0 \\ \frac{1}{y} & 0 \end{bmatrix}$. *Then*

$$\begin{aligned} ABA &= \begin{bmatrix} x & y \\ z & w \end{bmatrix}\begin{bmatrix} 0 & 0 \\ \frac{1}{y} & 0 \end{bmatrix}\begin{bmatrix} x & y \\ z & w \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ \frac{w}{y} & 0 \end{bmatrix}\begin{bmatrix} x & y \\ z & w \end{bmatrix} \\ &= \begin{bmatrix} x & y \\ \frac{wx}{y} & w \end{bmatrix} = \begin{bmatrix} x & y \\ z & w \end{bmatrix}. \end{aligned}$$

Similarly, if $z \neq 0$ *or* $w \neq 0$, *then we can find* B *such that* $ABA = A$. *Thus,* $M_2(\mathbb{R})$ *is a regular ring.*

Because $M_2(\mathbb{R})$ is not a division ring, it follows that a regular ring need not be a division ring. However, a division ring is a regular ring as shown in Example 7.2.5. In the next theorem, we show that a regular ring under a suitable condition becomes a division ring.

Theorem 7.2.8 *Let* R *be a regular ring with more than one element. Suppose for all* $x \in R$, *there exists a unique* $y \in R$ *such that* $x = xyx$. *Then*

(i) R *has no zero divisors,*

(ii) if $x \neq 0$ *and* $x = xyx$, *then* $y = yxy$ *for all* $x, y \in R$,

(iii) R *has an identity,*

(iv) R *is a division ring.*

Proof. (i) Let x be a nonzero element of R and $xz = 0$ for some $z \in R$. Now by the hypothesis, there exists a unique $y \in R$ such that $xyx = x$. Thus,

$$x(y - z)x = xyx - xzx = xyx.$$

Hence, by the uniqueness of y, $y - z = y$, so $z = 0$. This proves that R has no zero divisors.

(ii) Let $x \neq 0$ and $xyx = x$. Then

$$x(y - yxy) = xy - xyxy = xy - xy = 0.$$

Because R has no zero divisors and $x \neq 0$, $y - yxy = 0$, so $yxy = y$.

(iii) Let $0 \neq x \in R$. Then there exists a unique $y \in R$ such that $xyx = x$. Let $e = yx$. If $e = 0$, then $x = xyx = 0$, which is a contradiction. Therefore, $e \neq 0$. Also,

$$e^2 = yxyx = y(xyx) = yx = e.$$

Let $z \in R$. Then

$$(ze - z)e = ze^2 - ze = ze - ze = 0.$$

Thus, by (i), either $ze - z = 0$ or $ze = z$. Similarly, $e(ez - z) = 0$ implies that $ez = z$. Hence, e is the identity of R.

(iv) By (iii), R contains an identity element e. To show R is a division ring, it remains to be shown that every nonzero element of R has an inverse in R. Let x be a nonzero element in R. Then there exists a unique $y \in R$ such that $xyx = x$. Thus, $xyx = xe$, i.e., $x(yx - e) = 0$. Because R has no zero divisors and $x \neq 0$, $yx - e = 0$, so $yx = e$. Similarly, $xyx = ex$ implies $xy = e$. Therefore, $xy = e = yx$. Hence, R is a division ring. ■

Exercises

1. Prove that a Boolean ring R is a field if and only if R contains only 0 and 1.
2. Prove that a ring R with 1 is a Boolean ring if and only if for all $a, b \in R$, $(a + b)ab = 0$.
3. Let R be a Boolean ring with more than two elements. Find all zero divisors of R.
4. Let $T = \{f \mid f : \mathbb{R} \to \mathbb{Z}_2\}$. Define $+$ and $\cdot$ on T by for all $f, g \in T$, $(f + g)(x) = f(x) + g(x)$ and $(fg)(x) = f(x)g(x)$ for all $x \in \mathbb{R}$. Show that $(T, +, \cdot)$ is a Boolean ring.
5. Prove that a nonzero element of a regular ring with 1 is either a unit or a zero divisor.
6. Prove that the center of a regular ring is regular.
7. Let R be a ring in which each element is idempotent. Let $\overline{R} = R \times \mathbb{Z}_2$. Define $+$ and $\cdot$ on $\overline{R}$ by $(a, [n]) + (b, [m]) = (a + b, [n + m])$ and $(a, [n]) \cdot (b, [m]) = (na + mb + ab, [nm])$ for all $(a, [n]), (b, [m]) \in \overline{R}$. Show that $+$ and $\cdot$ are well defined on $\overline{R}$ and $\overline{R}$ is a Boolean ring.

8. Let R be a regular ring with 1.
   (i) Prove that for any $a \in R$, there exists an idempotent $e \in R$ such that $Ra = Re$.
   (ii) Prove that for any two idempotents $e, f \in R$, there exists an idempotent $g \in R$ such that $Re + Rf = Rg$.

Chapter 8

Subrings, Ideals, and Homomorphisms

The most important substructure of a ring is a particular subset called an "ideal." The term ideal was coined by Dedekind in honor of Kummer's work on ideal numbers. This notion of Kummer and Dedekind was used to obtain unique factorization properties. Kummer introduced the idea of an ideal number in his work on Fermat's last theorem. Noether followed with some important results on the theory of ideals. Some of her ideas were inspired by the work not only of Dedekind, but also of Kronecker and Lasker.

8.1 Subrings and Subfields

In this section, we introduce the idea of a subring of a ring. This concept is analogous to the concept of a subgroup of a group.

Definition 8.1.1 *Let $(R, +, \cdot)$ be a ring. Let R' be a subset of R. Then $(R', +, \cdot)$ is called a* ***subring*** *of $(R, +, \cdot)$ if*

(i) $(R', +)$ is a subgroup of $(R, +)$ and

(ii) for all $x, y \in R'$, $x \cdot y \in R'$.

Let $(R', +, \cdot)$ be a subring of the ring $(R, +, \cdot)$. Because $R' \subseteq R$ and because the associativity for $\cdot$ and the distributive laws are inherited, $(R', +, \cdot)$ is itself a ring. We will usually suppress the operations $+$ and $\cdot$ and call R' a subring of R. When R' and R are fields, R' is called a **subfield** of R.

The following theorem gives a necessary and sufficient condition for a subset to be a subring. With these conditions it is easy to verify whether a nonempty subset of a ring is a subring or not.

Theorem 8.1.2 *Let R be a ring. A nonempty subset R' of R is a subring of R if and only if $x - y \in R'$ and $xy \in R'$ for all $x, y \in R'$.*

Proof. First suppose that R' is a subring of R. Then R' is a ring. Hence, for all $x, y \in R$, $x - y$, $xy \in R'$.

Conversely, suppose $x - y \in R'$ and $xy \in R'$ for all $x, y \in R'$. Because $x - y \in R'$ for all $x, y \in R'$, $(R', +)$ is a subgroup of $(R, +)$ by Theorem 4.1.6. By the hypothesis, $xy \in R'$ for all $x, y \in R'$. Hence, R' is a subring of R. ■

Example 8.1.3 *(i) The ring $\mathbb{E}$ of even integers is a subring of $\mathbb{Z}$. $\mathbb{E}$ is without* 1.

(ii) Consider the subset $\mathbb{E}_8 = \{[0], [2], [4], [6]\}$ of $\mathbb{Z}_8$. Then $\mathbb{E}_8$ is a subring of $\mathbb{Z}_8$. Hence, $\mathbb{E}_8$ is commutative. However, $\mathbb{E}_8$ has no identity and $\mathbb{E}_8$ does have zero divisors, namely, $[2], [4]$, *and* $[6]$.

Example 8.1.4 *Let $Q_{\mathbb{Z}} = \{(a_1, a_2, a_3, a_4) \mid a_i \in \mathbb{Z},\ i = 1, 2, 3, 4\}$. Define $+$ and $\cdot$ on $Q_{\mathbb{Z}}$ as in Example* 7.1.26. *Because the difference and product of integers is an integer, we have*

$$(a_1, a_2, a_3, a_4) - (b_1, b_2, b_3, b_4) \in Q_{\mathbb{Z}}$$

and

$$(a_1, a_2, a_3, a_4) \cdot (b_1, b_2, b_3, b_4) \in Q_{\mathbb{Z}}$$

for all $(a_1, a_2, a_3, a_4), (b_1, b_2, b_3, b_4) \in Q_{\mathbb{Z}}$. *Hence,* $Q_{\mathbb{Z}}$ *is a subring of* $Q_{\mathbb{R}}$.

We note that $Q_{\mathbb{Z}}$ *is noncommutative, has an identity, and is without zero divisors. Now* $(0, 2, 0, 0) \in Q_{\mathbb{Z}}$ *and as in Example* $(0, 2, 0, 0)^{-1} = (0, -\frac{1}{2}, 0, 0)$. *However,* $(0, -\frac{1}{2}, 0, 0) \notin Q_{\mathbb{Z}}$. *Thus,* $Q_{\mathbb{Z}}$ *is not a division ring.*

Example 8.1.5 *Set* $Q_{\mathbb{E}} = \{(a_1, a_2, a_3, a_4) \mid a_i \in \mathbb{E},\ i = 1, 2, 3, 4\}$. *Define* $+$ *and* $\cdot$ *on* $Q_{\mathbb{E}}$ *as in Example* 7.1.26. *Because the difference and product of even integers is an even integer,*

$$(a_1, a_2, a_3, a_4) - (b_1, b_2, b_3, b_4) \in Q_{\mathbb{E}}$$

and

$$(a_1, a_2, a_3, a_4) \cdot (b_1, b_2, b_3, b_4) \in Q_{\mathbb{E}}$$

for all $(a_1, a_2, a_3, a_4), (b_1, b_2, b_3, b_4) \in Q_{\mathbb{E}}$. *It follows that* $Q_{\mathbb{E}}$ *is a subring of* $Q_{\mathbb{Z}}$. *In fact,* $Q_{\mathbb{E}}$ *is a noncommutative ring without identity and without zero divisors.*

Example 8.1.6 *Consider the ring* $M_2(\mathbb{Z})$ *of Example* 7.1.9. *Let* $M_2(\mathbb{E})$ *denote the set of all* 2×2 *matrices with entries from* $\mathbb{E}$. *Because the sum, difference, and product of even integers is an even integer, it follows that* $M_2(\mathbb{E})$ *is a subring of* $M_2(\mathbb{Z})$. *Also,* $M_2(\mathbb{E})$ *is a noncommutative ring without identity and with zero divisors.*

Following along the lines of Theorem 8.1.2, we can prove the next theorem. We leave its proof as an exercise.

Theorem 8.1.7 *Let* F *be a field. A nonempty subset* S *of* F *is a subfield of* F *if and only if*
(i) S *contains more than one element,*
(ii) $x - y, xy \in S$ *for all* $x, y \in S$, *and*
(iii) $x^{-1} \in S$ *for all* $x \in S$, $x \neq 0$. ■

Example 8.1.8 $\mathbb{Q}$ *and* $\mathbb{Q}[\sqrt{2}] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$ *are subfields of* $\mathbb{R}$ *(see Worked-Out Exercise* 4 *below).*

Theorem 8.1.9 *Let* R *be a ring (field) and* $\{R_i \mid i \in \Lambda\}$ *be a nonempty family of subrings (subfields) of* R. *Then* $\cap_{i \in \Lambda} R_i$ *is a subring (subfield) of* R.

Proof. Suppose that $\{R_i \mid i \in \Lambda\}$ be a nonempty family of subrings of R. Then $0 \in R_i$ for all $i \in \Lambda$. This implies that $0 \in \cap_{i \in \Lambda} R_i$, so $\cap_{i \in \Lambda} R_i \neq \emptyset$. Let $x, y \in \cap_{i \in \Lambda} R_i$. Then $x, y \in R_i$ for all $i \in \Lambda$. Because each R_i is a subring, $x - y, xy \in R_i$ for all $i \in \Lambda$. Hence, $x - y, xy \in \cap_{i \in \Lambda} R_i$. Hence, by Theorem 8.1.2, $\cap_{i \in \Lambda} R_i$ is a subring of R.

Similarly, we can show that if each R_i is a subfield of the field R, then $\cap_{i \in \Lambda} R_i$ is a subfield of R. ■

Remark 8.1.10 *It is interesting to note that the intersection of all subfields of* $\mathbb{R}$ *is* $\mathbb{Q}$.

Worked-Out Exercises

♢ **Exercise 1** Let X be an infinite set. Then $(\mathcal{P}(X), \Delta, \cap)$ is a ring with 1. Let

$$R = \{A \in \mathcal{P}(X) \mid A \text{ is finite}\}.$$

Prove the following assertions.

(a) R is a subring of $\mathcal{P}(X)$.

(b) R is without identity.

(c) For all $A \in R$, $A \neq \emptyset$, A is a zero divisor in R.

(d) For all $A \in \mathcal{P}(X)$, $A \neq X$, $A \neq \emptyset$, A is a zero divisor in $\mathcal{P}(X)$.

Solution: (a) Because $\emptyset$ is finite, $\emptyset \in R$, so R is nonempty. Let $A, B \in R$. Then A and B are finite, so $A \cap B$ is finite. Now $A \Delta B = (A \cup B) \backslash (A \cap B)$, so $A \Delta B$ is finite. Therefore, $A \Delta B$, $A \cap B \in R$. Thus, R is closed under the operations Δ and $\cap$. Now it is easy to verify that $(R, \Delta, \cap)$ is a subring.

(b) Suppose R has an identity, say, E. Then E is finite. Because X is infinite, there exists $a \in X$ such that $a \notin E$. Now $\{a\} \in R$. Thus, $\{a\} = E \cap \{a\} = \emptyset$, which is a contradiction. Hence, R has no identity.

(c) Let $A \in R$ and $A \neq \emptyset$. Because A is finite and X is infinite, there exists $x \in X$ such that $x \notin A$. Now $\{x\} \in R$. Because $A \cap \{x\} = \emptyset$, A is a zero divisor.

(d) Let $A \in \mathcal{P}(X)$ be such that $A \neq X$ and $A \neq \emptyset$. Then there exists $x \in X$ such that $x \notin A$. Hence, $A \cap \{x\} = \emptyset$, so A is a zero divisor.

Exercise 2 Let R be a ring such that $a^2 + a$ is in the center of R for all $a \in R$. Show that R is commutative.

Solution: Let $x, y \in R$. Then $(x+y)^2+x+y \in C(R)$, i.e., $x^2+xy+yx+y^2+x+y \in C(R)$. Because $x^2+x, y^2+y \in C(R)$ and $C(R)$ is a subring (Exercise 14, page 148), $xy + yx \in C(R)$. Therefore, $x(xy + yx) = (xy + yx)x$, so $x^2y + xyx = xyx + yx^2$. Thus, $x^2y = yx^2$. Now $x^2 + x \in C(R)$, so $y(x^2 + x) = (x^2 + x)y$. Hence, $yx^2 + yx = x^2y + xy$, so $xy = yx$, proving that R is commutative.

♢ **Exercise 3** Find all subrings of the ring $\mathbb{Z}$ of integers. Find those subrings which do not contain the identity element.

Solution: Let n be a nonnegative integer and $T_n = n\mathbb{Z} = \{nt \mid t \in \mathbb{Z}\}$. Because $0 \in T_n$, $T_n \neq \emptyset$. Let $a = nt$, $b = ns$ be two elements in T_n. Then $a - b = nt - ns = n(t - s) \in T_n$ and $ab = (nt)(ns) = n(t(ns)) \in T_n$. Hence, T_n is a subring of $\mathbb{Z}$. We now show that if A is any subring of $\mathbb{Z}$, then $A = T_n$ for some nonnegative integer n.

Let A be a subring of $\mathbb{Z}$. If $A = \{0\}$, then $A = 0\mathbb{Z}$. Suppose $A \neq \{0\}$. Then there exists $m \in A$ such that $m \neq 0$. Now $-m \in A$, so A contains a positive integer. By the well-ordering principle, A contains a smallest positive integer. Let n be the smallest positive integer in A. Then $n\mathbb{Z} \subseteq A$. Let $m \in A$. By the division algorithm, there exist integers q and r such that $m = nq + r$, $0 \leq r < n$. Because $n \in A$, $nq \in A$. Hence, $r = m - nq \in A$. The minimality of n implies that $r = 0$, so $m = nq \in n\mathbb{Z}$. Thus, $A = n\mathbb{Z}$. If $n \neq 1$, then $n\mathbb{Z}$ does not contain identity.

♢ **Exercise 4** Show that $\mathbb{Q}[\sqrt{2}] = \{a + b\sqrt{2} \in \mathbb{R} \mid a, b \in \mathbb{Q}\}$ is a subfield of the field $\mathbb{R}$.

Solution: Because $0 = 0 + 0\sqrt{2} \in \mathbb{Q}[\sqrt{2}]$, $\mathbb{Q}[\sqrt{2}] \neq \emptyset$. Let $a + b\sqrt{2}$, $c + d\sqrt{2} \in \mathbb{Q}[\sqrt{2}]$. Then

$$(a + b\sqrt{2}) - (c + d\sqrt{2}) = (a - c) + (b - d)\sqrt{2} \in \mathbb{Q}[\sqrt{2}]$$

and

$$(a + b\sqrt{2})\ (c + d\sqrt{2}) = (ac + 2bd) + (ad + bc)\sqrt{2} \in \mathbb{Q}[\sqrt{2}].$$

Now $0 + 0\sqrt{2}$ and $1 + 0\sqrt{2}$ are distinct elements of $\mathbb{Q}[\sqrt{2}]$. Therefore, $\mathbb{Q}[\sqrt{2}]$ contains more than one element. Let $a + b\sqrt{2}$ be a nonzero element of $\mathbb{Q}[\sqrt{2}]$. Then a and b cannot both be zero simultaneously. We now show that $a - b\sqrt{2} \neq 0$. Suppose $a - b\sqrt{2} = 0$. Then $a = b\sqrt{2}$. If $b = 0$, then $a = 0$. Therefore, both a and b are zero, a contradiction. If $b \neq 0$, then $\sqrt{2} = \frac{a}{b} \in \mathbb{Q}$, a contradiction. Hence, $a - b\sqrt{2} \neq 0$. Similarly, $a + b\sqrt{2} \neq 0$. Thus, $a^2 - 2b^2 = (a + b\sqrt{2})(a - b\sqrt{2}) \neq 0$. Now

$$\frac{1}{a + b\sqrt{2}} = \frac{(a - b\sqrt{2})}{a^2 - 2b^2} = \frac{a}{a^2 - 2b^2} - \frac{b}{a^2 - 2b^2}\sqrt{2} \in \mathbb{Q}[\sqrt{2}].$$

Because $(a + b\sqrt{2})(\frac{1}{a+b\sqrt{2}}) = 1$, $(a + b\sqrt{2})^{-1}$ exists in $\mathbb{Q}[\sqrt{2}]$. Thus, we find that $\mathbb{Q}[\sqrt{2}]$ is a subfield of $\mathbb{R}$ by Theorem 8.1.7.

Exercises

1. Prove the following the statements.
   (i) $T_1 = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} \mid a, b, c \in \mathbb{Z} \right\}$ is a subring of $M_2(\mathbb{Z})$.
   (ii) $T_2 = \left\{ \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \mid a, b \in \mathbb{Z} \right\}$ is a subring of $M_2(\mathbb{Z})$.
   (iii) $T_3 = \left\{ \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} \mid a \in \mathbb{Z} \right\}$ is a subring of $M_2(\mathbb{Z})$.
   (iv) $T_4 = \left\{ \begin{bmatrix} a & b \\ 0 & a \end{bmatrix} \mid a, b \in \mathbb{Z} \right\}$ is a subring of T_1.
2. In the ring $\mathbb{Z}$ of integers, find which of the following subsets of $\mathbb{Z}$ are subrings.
   (i) The set of integers of the form $4k + 2$, $k \in \mathbb{Z}$.
   (ii) The set of integers of the form $4k + 1$, $k \in \mathbb{Z}$.
   (iii) The set of integers of the form $4k$, $k \in \mathbb{Z}$.
3. Show that $T = \{[0], [5]\}$ is a subring of the ring $\mathbb{Z}_{10}$.
4. Let R be a ring with 1. Show that the subset $T = \{n1 \mid n \in \mathbb{Z}\}$ is a subring of R.
5. Let R be a ring and n be a positive integer. Show that the subset $T = \{a \in R \mid na = 0\}$ is a subring of R.

6. Show that $T = \left\{ \begin{bmatrix} a & b\sqrt{3} \\ -b\sqrt{3} & a \end{bmatrix} \mid a, b \in \mathbb{R} \right\}$ is a subring of $M_2(\mathbb{R})$.
7. Show that $\mathbb{Q}[\sqrt{3}]$ and $\mathbb{Q}[\sqrt{5}]$ are subfields of the field $\mathbb{R}$, but $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Z}\}$ is not a subfield of $\mathbb{R}$.
8. Show that $\mathbb{Q}(i) = \{a + bi \mid a, b \in \mathbb{Q}\}$ is a subfield of $\mathbb{C}$, where $i^2 = -1$.
9. Show that $F = \left\{ \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \mid a, b \in \mathbb{Z}_5 \right\}$ is a subring of $M_2(\mathbb{Z}_5)$. Is F a field?
10. Let ω be a root of $x^2 + x + 1 = 0$. Prove that $T = \{a + b\omega \mid a, b \in \mathbb{Q}\}$ is a subfield of the field of complex numbers.
11. Let F be a field of characteristic $p > 0$. Show that $T = \{a \in F \mid a^p = a\}$ is a subfield of F.
12. Prove that $T = \left\{ \begin{bmatrix} x + y & y \\ -y & x \end{bmatrix} \mid x, y \in \mathbb{Z} \right\}$ is a subring of $M_2(\mathbb{Z})$. Also, show that every nonzero element of T is a unit in $M_2(\mathbb{R})$.
13. Let R be a commutative ring. Show that the set
$$T = \{r \in R \mid r^n = 0 \text{ for some integer } n\}$$
is a subring of R.
14. Prove that $C(R)$ is a subring of R and that $C(R)$ is commutative.
15. Let e be an idempotent of a ring R. Prove that the set
$$eRe = \{ere \mid r \in R\}$$
is a subring of R with e as the identity element.
16. Find the center of the ring $M_2(\mathbb{R})$.
17. Prove that the characteristic of a subfield is the same as the characteristic of the field.
18. Find all subrings with identity of the ring $\mathbb{Z}_{16}$.
19. Find all subfields of the field $\mathbb{Z}_p$, p a prime integer.
20. Let R be a ring without any nonzero nilpotent elements. Show that $(ara - ra)^2 = 0$ for all $r \in R$ and for all idempotent elements $a \in R$. Hence, show that $C(R)$ contains all idempotent elements.
21. Let $C = \{ f : \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous on } \mathbb{R}\}$. Define $+$ and $\cdot$ on C by
$$\begin{aligned} (f + g)(x) &= f(x) + g(x), \\ (f \cdot g)(x) &= f(x)g(x) \end{aligned}$$
for all $f, g \in C$ and for all $x \in \mathbb{R}$.
(i) Show that C is a ring.
(ii) Let $D = \{f \in C \mid f \text{ is differentiable on } \mathbb{R}\}$. Show that D is a subring of C.
22. Let R be a ring and $f : R \to [0, 1]$ be such that
$$\begin{aligned} f(a - b) &\geq \min\{f(a), f(b)\}, \\ f(ab) &\geq \min\{f(a), f(b)\} \end{aligned}$$
for all $a, b \in R$. Prove that for all $t \in \mathcal{I}(f)$, $R_t = \{x \in R \mid f(x) \geq t\}$ is a subring of R.
23. In the following exercises, write the proof if the statement is true; otherwise, give a counterexample.
(i) The union of two subrings of a ring is a subring.
(ii) The identity element of a subring is always the identity element of the ring.
(iii) $\mathbb{Q}$ is the only subfield of the field $\mathbb{R}$.
(iv) $\mathbb{Q}[\sqrt{3}] = \{a + b\sqrt{3} \mid a, b \in \mathbb{Q}\}$ is the intersection of all subfields of $\mathbb{R}$ which contain $\sqrt{3}$.
(v) The set $\mathbb{Z}$ of integers is a subring of the field of real numbers.
(vi) Every additive subgroup of $\mathbb{Z}$ is a subring of $\mathbb{Z}$.

8.2 Ideals and Quotient Rings

In this section, we introduce the notions of ideals and quotient rings. These concepts are analogous to normal subgroups and quotient groups.

The very famous problem called "Fermat's last theorem" led to the invention of ideals. Fermat (1601–1665) jotted many of his results in the margin of *Diophantus' Arithmetica*. For this particular "theorem," Fermat wrote that he discovered a remarkable theorem whose proof was too long to put in the margin. The theorem is stated as follows: If n is an integer greater than 2, then there exist no positive integers x, y, z such that $x^n + y^n = z^n$. However, no one was able to prove this result until recently; in 1994, Andrew Wiles found a proof after many years of work.

In 1843, Kummer (1810–1893) thought that he had found a proof of Fermat's last theorem. However, Kummer had incorrectly assumed uniqueness of the factorization of complex numbers of the form $x+\lambda y$, where $\lambda^p = 1$ for p an odd prime. Dirichlet (1805–1859) had made an incorrect assumption about factorization of numbers. Kummer continued his efforts to solve Fermat's last theorem. He was partially successful by introducing the concept of "ideal number." Dedekind (1831–1916) used Kummer's ideas to invent the notion of an ideal. Kronecker (1823–1891) also played an important part in the development of ring theory.

Definition 8.2.1 *Let R be a ring. Let I be a nonempty subset of R.*
*(i) I is called a **left ideal** of R if for all $a, b \in I$ and for all $r \in R$, $a - b \in I$, $ra \in I$.*
*(ii) I is called a **right ideal** of R if for all $a, b \in I$ and for all $r \in R$, $a - b \in I$, $ar \in I$.*
*(iii) I is called a (**two-sided**) **ideal** of R if I is both a left and a right ideal of R.*

From the definition of a left (right) ideal, it follows that if I is a left (right) ideal of R, then I is a subring of R. Also, if R is a commutative ring, then every left ideal is also a right ideal and every right ideal is a left ideal. Thus, for commutative rings every left or right ideal is an ideal.

By Theorem 8.1.2, it follows that a nonempty subset I of a ring R is an ideal if and only if $(I, +)$ is a subgroup of $(R, +)$ and for all $a \in I$ and for all $r \in R$, ar and $ra \in I$.

Example 8.2.2 *Let R be a ring. The subsets $\{0\}$ and R of R are (left, right) ideals. These ideals are called **trivial** ideals. All other (left, right) ideals are called **nontrivial.***

An ideal I of a ring R is called a **proper** ideal if $I \neq R$.

Example 8.2.3 *Let $n \in \mathbb{Z}$ and $I = \{nk \mid k \in \mathbb{Z}\}$. As in Worked-Out Exercise 3 (page 147), I is a subring. Also, for all $r \in \mathbb{Z}$, $(nk)r = n(kr) \in I$ and $r(nk) = n(rk) \in I$. Hence, I is an ideal of $\mathbb{Z}$.*

Next, we give an example of a ring in which there exists a left ideal which is not a right ideal, a right ideal which is not a left ideal, and a subring which is not a left (right) ideal.

Example 8.2.4 *Consider the ring $M_2(\mathbb{Z})$. Let*

$$I_1 = \left\{ \begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix} \mid a, b \in \mathbb{Z} \right\},$$

$$I_2 = \left\{ \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix} \mid a, b \in \mathbb{Z} \right\},$$

$$I_3 = \left\{ \begin{bmatrix} a & c \\ b & d \end{bmatrix} \mid a, b, c \text{ and } d \text{ are even integers} \right\},$$

and

$$I_4 = \left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \mid a \in \mathbb{Z} \right\}.$$

Because $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \in I_1$, $I_1 \neq \emptyset$. *Let* $\begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix}, \begin{bmatrix} c & 0 \\ d & 0 \end{bmatrix} \in I_1$ *and* $\begin{bmatrix} x & y \\ z & w \end{bmatrix} \in M_2(\mathbb{Z})$. *Then*

$$\begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix} - \begin{bmatrix} c & 0 \\ d & 0 \end{bmatrix} = \begin{bmatrix} a-c & 0 \\ b-d & 0 \end{bmatrix} \in I_1$$

and

$$\begin{bmatrix} x & y \\ z & w \end{bmatrix} \begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix} = \begin{bmatrix} xa+yb & 0 \\ za+wb & 0 \end{bmatrix} \in I_1,$$

proving that I_1 is a left ideal of $M_2(\mathbb{Z})$. Now $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \in I_1$ *and* $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \in M_2(\mathbb{Z})$, *but*

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \notin I_1.$$

Hence, I_1 is not a right ideal of $M_2(\mathbb{Z})$. Similarly, I_2 is a right ideal of $M_2(\mathbb{Z})$, but not a left ideal, I_3 is an ideal of $M_2(\mathbb{Z})$, and I_4 is a subring, but not an ideal of $M_2(\mathbb{Z})$.

We remind the reader to notice the similarity of the next few results with corresponding results in linear algebra and group theory.

Theorem 8.2.5 *Let R be a ring and $\{I_\alpha \mid \alpha \in \Lambda\}$ be a nonempty collection of left (right) ideals of R. Then $\cap_{\alpha\in\Lambda} I_\alpha$ is a left (right) ideal of R.*

Proof. Suppose $\{I_\alpha \mid \alpha \in \Lambda\}$ is nonempty a collection of left ideals of R. Because $0 \in I_\alpha$ for all α, $0 \in \cap_\alpha I_\alpha$. Thus, $\cap_\alpha I_\alpha \neq \emptyset$. Let $a, b \in \cap_\alpha I_\alpha$. Then $a, b \in I_\alpha$ for all α. Because each I_α is a left ideal, $a - b \in I_\alpha$ for all α. Hence, $a - b \in \cap_\alpha I_\alpha$. Let $r \in R$. Because each I_α is a left ideal of R, $ra \in I_\alpha$ for all α. This implies that $ra \in \cap_\alpha I_\alpha$. Thus, $\cap_\alpha I_\alpha$ is a left ideal of R.

Similarly, if $\{I_\alpha \mid \alpha \in \Lambda\}$ is a nonempty collection of right ideals of R, then $\cap_\alpha I_\alpha$ is a right ideal of R. ■

Notation 8.2.6 *Let $a_1, a_2, \ldots, a_n \in R$. Then by the notation $\sum_{i=1}^n a_i$, we mean the sum $a_1 + a_2 + \cdots + a_n$.*

Definition 8.2.7 *Let S be a nonempty subset of a ring R.*

(i) Define $\langle S\rangle_l$ to be the intersection of all left ideals of R which contain S.

(ii) Define $\langle S\rangle_r$ to be the intersection of all right ideals of R which contain S.

(iii) Define $\langle S\rangle$ to be the intersection of all right ideals of R which contain S.

Using Theorem 8.2.5, the following theorem is immediate.

Theorem 8.2.8 *Let S be a nonempty subset of a ring R.*

(i) $\langle S\rangle_l$ is a left ideal. The left ideal $\langle S\rangle_l$ is called the ***left ideal generated*** *by S.*

(ii) $\langle S\rangle_r$ is a right ideal of R. The right ideal $\langle S\rangle_r$ is called the ***right ideal generated*** *by S.*

(iii) $\langle S\rangle$ is an ideal of R. The ideal $\langle S\rangle$ is called the ***ideal generated*** *by S.*

Note that $\langle S\rangle_l$ is the smallest left ideal of R which contains S.

Theorem 8.2.9 *Let R be a ring and S be a nonempty subset of R. Then*

(i)

$$\begin{aligned}\langle S\rangle_l &= \{\textstyle\sum_{i=1}^k r_i s_i + \sum_{j=1}^l n_j t_j \mid r_i \in R, n_j \in \mathbb{Z}, s_i, t_j \in S,\\ &\quad 1 \le i \le k, 1 \le j \le l, k, l \in \mathbb{N}\}.\end{aligned}$$

(ii)

$$\begin{aligned}\langle S\rangle_r &= \{\textstyle\sum_{i=1}^k s_i r_i + \sum_{j=1}^l n_j t_j \mid r_i \in R, n_j \in \mathbb{Z}, s_i, t_j \in S,\\ &\quad 1 \le i \le k, 1 \le j \le l, k, l \in \mathbb{N}\}.\end{aligned}$$

Proof. (i) Let

$$\begin{aligned}A &= \{\textstyle\sum_{i=1}^k r_i s_i + \sum_{j=1}^l n_j t_j \mid r_i \in R, n_j \in \mathbb{Z}, s_i, t_j \in S,\\ &\quad 1 \le i \le k, 1 \le j \le l, k, l \in \mathbb{N}\}.\end{aligned}$$

Because $\langle S\rangle_l$ is the intersection of all left ideals of R which contain S, it follows that $\langle S\rangle_l \supseteq S$. Let $\sum_{i=1}^k r_i s_i + \sum_{j=1}^l n_j t_j \in A$, where $r_i \in R, n_j \in \mathbb{Z}, s_i, t_j \in S$, $1 \le i \le k, 1 \le j \le l$, $k, l \in \mathbb{N}$. Now $s_i, t_j \in S \subseteq \langle S\rangle_l$. Because $\langle S\rangle_l$ is a left ideal, $r_i s_i$, $n_j t_j \in \langle S\rangle_l$, $1 \le i \le k, 1 \le j \le l$. Once again using the fact that $\langle S\rangle_l$ is a left ideal, we can conclude that $\sum_{i=1}^k r_i s_i + \sum_{j=1}^l n_j t_j \in \langle S\rangle_l$. Hence, $A \subseteq \langle S\rangle_l$.

We now show that A is a left ideal of R such that $S \subseteq A$. We can then conclude that $\langle S\rangle_l \subseteq A$ because $\langle S\rangle_l$ is the smallest left ideal of R containing S.

Let $s \in S$. Then $s = 0 \cdot s + 1s \in A$. Hence, $S \subseteq A$. Let $\sum_{i=1}^k r_i s_i + \sum_{j=1}^l n_j t_j$ and $\sum_{i=1}^p r'_i\, s'_i + \sum_{j=1}^q n'_j t'_j \in A$. Then

$$\begin{aligned}&\left(\textstyle\sum_{i=1}^k r_i s_i + \sum_{j=1}^l n_j t_j\right) - \left(\sum_{i=1}^p r'_i s'_i + \sum_{j=1}^q n'_j t'_j\right)\\ = \;&\left(\textstyle\sum_{i=1}^k r_i s_i + \sum_{i=1}^p (-r'_i) s'_i\right) + \left(\sum_{j=1}^l n_j t_j + \sum_{j=1}^q (-n'_j) t'_j\right) \in A\end{aligned}$$

Let $r \in R$. Then $r\left(\sum_{i=1}^{k} r_i s_i + \sum_{j=1}^{l} n_j t_j\right) = \sum_{i=1}^{k}(rr_i)s_i + \sum_{j=1}^{l}(n_j r)t_j \in A$. Thus, A is a left ideal of R. It now follows that $\langle S\rangle_l \subseteq A$.

Consequently, $\langle S\rangle_l = A$

(ii) The proof is similar to (i). ■

Corollary 8.2.10 *Let R be a ring and S be a nonempty subset of R. If R is with 1, then*

(i)

$$\langle S\rangle_l = \left\{\sum_{i=1}^{k} r_i s_i \mid r_i \in R, s_i \in S, 1 \le i \le k,\ n \in \mathbb{N}\right\}.$$

(ii)

$$\langle S\rangle_r = \left\{\sum_{i=1}^{k} s_i r_i \mid r_i \in R, s_i \in S, 1 \le i \le k,\ n \in \mathbb{N}\right\}.$$

Proof. (i) Let $A = \left\{\sum_{i=1}^{k} r_i s_i \mid r_i \in R, s_i \in S, 1 \le i \le k,\ n \in \mathbb{N}\right\}$. Observe that $A \subseteq \langle S\rangle_l$.

Let $\sum_{i=1}^{t} r_i s_i + \sum_{j=1}^{l} n_j t_j \in \langle S\rangle_l$. Because R has an identity 1, $n_j t_j = (n_j 1)t_j$ and $n_j 1 \in R$. Thus, $\sum_{i=1}^{t} r_i s_i + \sum_{j=1}^{l} n_j t_j = \sum_{i=1}^{t} r_i s_i + \sum_{j=1}^{l}(n_j 1)t_j \in A$. Hence, $\langle S\rangle_l \subseteq A$. Consequently,

$$\langle S\rangle_l = \left\{\sum_{i=1}^{k} r_i s_i \mid r_i \in R, s_i \in S, 1 \le i \le k,\ n \in \mathbb{N}\right\}.$$

(ii) The proof is similar to (i). ■

Let us note the following:

1. If $S = \{a_1, a_2, \ldots, a_n\}$, then the left ideal $\langle S\rangle_l$ generated by S is denoted by $\langle a_1, a_2, \ldots, a_n\rangle_l$. In this case, we call $\langle S\rangle_l$ a **finitely generated left ideal**. Similar terminology is used for $\langle S\rangle_r$ and $\langle S\rangle$.
2. If $S = \{a\}$, then
   (a) $\langle a\rangle_l$ is called the **principal left ideal** generated by a,
   (b) $\langle a\rangle_r$ is called the **principal right ideal** generated by a, and
   (c) $\langle a\rangle$ is called the **principal ideal** generated by a.

Corollary 8.2.11 *Let R be a ring and $a \in R$.*

(i) Then

$$\langle a\rangle_l = \{ra + na \mid r \in R,\ n \in \mathbb{Z}\}.$$

(ii) If R is with 1, then

$$\langle a\rangle_l = \{ra \mid r \in R\}.$$

Proof. (i) This assertion follows from the equality

$$\sum_{i=1}^{k} r_i a + \sum_{j=1}^{m} n_j a = (\sum_{i=1}^{k} r_i)a + (\sum_{j=1}^{m} n_j)a.$$

(ii) This follows from (i) and Corollary 8.2.10. ■

Similarly, we can prove that $\langle a\rangle_r = \{ar + na \mid r \in R, n \in \mathbb{Z}\}$ and $\langle a\rangle = \{ra + as + na + \sum_{i=1}^{k} r_i a s_i \mid r, s, r_i, s_i \in R,\ n \in \mathbb{Z}, 1 \le i \le k,\ k \in \mathbb{N}\}$.

Consider the subsets $Ra = \{ra \mid r \in R\}$ and $aR = \{ar \mid r \in R\}$ of R. If R is without identity, then Ra (aR) is still a left (right) ideal of R (Exercise 4, page 157). It is not necessarily the case that $a \in Ra$ ($a \in aR$) as illustrated by the next example.

Example 8.2.12 *Consider the ring $\mathbb{E}$ of even integers. $\mathbb{E}$ does not have an identity. Now*

$$\langle 2\rangle = \{r2 + n2 \mid r \in \mathbb{E}, n \in \mathbb{Z}\} = \{0, \pm 2, \pm 4, \ldots\}$$

and $2 \in \langle 2\rangle$. However,

$$R2 = \{r2 \mid r \in \mathbb{E}\} = \{0, \pm 4, \pm 8, \ldots\},$$

which does not contain 2.

In the next theorem, we obtain a necessary and sufficient condition for a ring with 1 to be a division ring.

Theorem 8.2.13 *Let R be a ring with 1. Then R is a division ring if and only if R has no nontrivial left ideals.*

Proof. Suppose R is a division ring. Let I be a left ideal of R such that $I \supset \{0\}$. Then there exists $a \in I$ such that $a \neq 0$. Now $a \neq 0$, so $a^{-1} \in R$. Thus, because I is a left ideal, $1 = a^{-1}a \in I$. This implies that for all $r \in R$, $r = r1 \in I$. This shows that $R \subseteq I$. Because, $I \subseteq R$, we can now conlude that $R = I$. Consequently, R has no nontrivial left ideals.

Conversely, suppose R has no nontrivial left ideals. Let $a \in R$ and $a \neq 0$. We show that a is a unit.

Because R has no nontrivial left ideals and $a \neq 0$, we must have $\langle a \rangle_l = R$. This implies that $1 \in \langle a \rangle_l$. By Corollary 8.2.11(ii), $\langle a \rangle_l = \{ra \mid r \in R\}$. Thus, there exists $r \in R$ such that $1 = ra$. This implies that $r \neq 0$.

Now $r \neq 0$, so proceeding as in the case of the nonzero element a, we can show that $tr = 1$ for some $t \in R$. Thus, we have

$$t = t1 = t(ra) = (tr)a = 1a = a.$$

This, implies that $ar = 1$. Hence,

$$ra = 1 = ar,$$

i.e., a is a unit. Consequently, every nonzero element of R is a unit. Hence, R is a division ring. ■

Following along the lines of the above theorem, we can prove that a ring R with 1 is a division ring if and only if R has no nontrivial right ideals.

The following corollary is immediate from Theorem 8.2.13.

Corollary 8.2.14 *Let R be a commutative ring with 1. Then R is a field if and only if R has no nontrivial ideals.* ■

Definition 8.2.15 *A ring R is called a **simple** ring if $R^2 \neq \{0\}$ and $\{0\}$ and R are the only ideals of R.*

Example 8.2.16 *Every division ring is a simple ring.*

Example 8.2.17 *In this example, we show that $M_2(\mathbb{R})$ is a simple ring. Let A be a nonzero ideal of $M_2(\mathbb{R})$. Then there exists a nonzero element $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \in A$. Now at least one of a, b, c, d is nonzero. Because A is an ideal and $\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \in M_2(\mathbb{R})$, we have*

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} b & 0 \\ d & 0 \end{bmatrix} \in A,$$

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} c & d \\ 0 & 0 \end{bmatrix} \in A,$$

and

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} d & 0 \\ 0 & 0 \end{bmatrix} \in A.$$

Therefore, we find that A contains a matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ such that $a \neq 0$. Now $a^{-1} \in \mathbb{R}$ and

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} a^{-1} & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ ca^{-1} & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \in A.$$

Thus,

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

Finally,

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

Hence,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \in A.$$

This implies that $A = M_2(\mathbb{R})$. Also note that $M_2(\mathbb{R})$ is not a division ring.

Example 8.2.17 shows that there are simple rings, which are not division rings.

Notation 8.2.18 *For $a \in R$, aRa denotes the set $\{ara \mid r \in R\}$.*

We now consider the sum and product of left (right) ideals.

Let A and B be two nonempty subsets of a ring R. Define the **sum** and **product** of A and B as follows:

$$A + B = \{a + b \mid a \in A,\ b \in B\}$$

$$AB = \{a_1b_1 + a_2b_2 + \cdots + a_nb_n \mid a_i \in A,\ b_i \in B,\ i = 1, 2, \ldots, n,\ n \in \mathbb{N}\}.$$

Thus, AB denotes the set of all finite sums of the form $\sum a_ib_i$, $a_i \in A$, $b_i \in B$.

Let $n \in \mathbb{N}$. Inductively, we define

$$\begin{array}{rcll} A^1 & = & A, & \\ A^n & = & AA^{n-1} & \text{if } n > 1. \end{array}$$

Example 8.2.19 *Consider $\mathbb{Z}$, the ring of integers. Let $A = \langle 2 \rangle$, the ideal generated by 2, and $B = \langle 3 \rangle$, the ideal generated by 3. Note that $A = \{2n \mid n \in \mathbb{Z}\}$ and $B = \{3n \mid n \in \mathbb{Z}\}$. Let $m \in \mathbb{Z}$. Now $m = 2(-m) + 3m \in A + B$. This implies that $A + B = \mathbb{Z}$.*

Next we dsetermine AB. Let $m \in \mathbb{Z}$. Now $6n = 2 \cdot (3n) \in AB$. This implies that $\langle 6 \rangle = \{6n \mid n \in \mathbb{Z}\} \subseteq AB$. Let $x \in AB$. Then

$$x = a_1b_1 + a_2b_2 + \cdots + a_nb_n,$$

for some $n \in \mathbb{N}$, where $a_i \in A$, $b_i \in B$, $i = 1, 2, \ldots, n$. Now $a_i = 2t_i$ and $b_i = 3s_i$, for some $t_i, s_i \in \mathbb{Z}$, $i = 1, 2, \ldots, n$. Hence

$$\begin{array}{rcl} x & = & a_1b_1 + a_2b_2 + \cdots + a_nb_n \\ & = & (2t_1)(3s_1) + (2t_2)(3s_2) + \cdots + (2t_n)(3s_n) \\ & = & 6(t_1s_1) + 6(t_2s_2) + \cdots + 6(t_ns_n) \\ & = & 6(t_1s_1 + t_2s_2 + \cdots + t_ns_n) \\ & = & 6k \in \langle 6 \rangle, \text{ where } k = t_1s_1 + t_2s_2 + \cdots + t_ns_n \in \mathbb{Z}. \end{array}$$

Thus, $AB \subseteq \langle 6 \rangle$. Hence, $AB = \langle 6 \rangle$.

We now list some interesting properties of the sum and product of left (right) ideals.

Theorem 8.2.20 *Let A, B, and C be left (right) ideals of a ring R. Then the following assertions hold.*

(i) $A \subseteq A + B$.

(ii) $A + B = B + A$ is a left (right) ideal of R.

(iii) $A + A = A$.

(iv) $(A + B) + C = A + (B + C)$.

(v) AB is a left (right) ideal of R.

(vi) $(AB)C = A(BC)$.

(vii) If A, B and C are ideals, then $A(B + C) = AB + AC$, $(B + C)A = BA + CA$.

(viii) If A is a right ideal and B is a left ideal, then $AB \subseteq A \cap B$.

(ix) R is a regular ring if and only if for any right ideal A and for any left ideal B, $AB = A \cap B$.

Proof. We only prove (ix) and (x) and leave the other properties as exercises.

(ix) Suppose R is a regular ring. Let $a \in A \cap B$. There exists $b \in R$ such that $a = aba$. Because B is a left ideal and $a \in B$, $ba \in B$. Thus, $a = a(ba) \in AB$, whence $A \cap B \subseteq AB$. By (vii), $AB \subseteq A \cap B$. Consequently, $AB = A \cap B$. Conversely, assume that $AB = A \cap B$ for any right ideal A and left ideal B of R. Let $a \in R$ and consider $\langle a \rangle_r$, the right ideal generated by a. Because $\langle a \rangle_r$ is a right ideal, $\langle a \rangle_r R \subseteq \langle a \rangle_r$. Also, by our assumption $\langle a \rangle_r \cap R = \langle a \rangle_r R$. Hence,

$$a \in \langle a \rangle_r \cap R = \langle a \rangle_r R.$$

Therefore, $a = \sum_{i=1}^{n} a_ib_i$ for some $a_i \in \langle a \rangle_r$, $b_i \in R$, $i = 1, 2, \ldots, n$. From the statements following Corollary 8.2.11, $a_i = at_i + n_ia$ for some $t_i \in R$, $n_i \in \mathbb{Z}$, $i = 1, 2, \ldots, n$. Thus,

$$a = \sum_{i=1}^{n} a_ib_i = \sum_{i=1}^{n} (at_i + n_ia)b_i = a(\sum_{i=1}^{n} (t_ib_i + n_ib_i)) \in aR.$$

This implies that $\langle a \rangle_r = aR$. Because $aR \subseteq \langle a \rangle_r$, $\langle a \rangle_r = aR$. Similarly, $\langle a \rangle_l = Ra$. It now follows that $a \in aR \cap Ra = (aR)(Ra) \subseteq aRa$. Hence, there exists $b \in R$ such that $a = aba$, i.e., a is regular. Consequently, R is regular. ■

Quotient Rings

We now give the analogue of quotient groups for rings. Let R be a ring and I an ideal of R. Let $x \in R$. Let $x+I$ denote the set

$$x + I = \{x + a \mid a \in I\}.$$

The set $x + I$ is called a **coset** of I. For $x, y \in R$, we leave it as exercise for the reader to verify that

$$x + I = y + I \text{ if and only if } x - y \in I.$$

This property of cosets is, in fact, analogous to property of cosets for a group, (see Theorem 4.3.3). Moreover notice that

$$0 + I = I.$$

Let R/I denote the set

$$R/I = \{x + I \mid x \in R\}.$$

Because $I = 0 + I \in R/I$, R/I is a nonempty set. Define the operations $+$ and $\cdot$ on R/I as follows: for all $x + I$, $y + I \in R/I$

$$(x + I) + (y + I) = (x + y) + I,$$

and

$$(x + I) \cdot (y + I) = xy + I.$$

We leave it as an exercise for the reader to verify that $+$ and $\cdot$ are binary operations on R/I.

Under these binary operations $(R/I, +, \cdot)$ satiesfies the properties of a ring. Let us verify some of these properties.

Let $x + I$, $y + I$, $z + I \in R/I$. Now

$$\begin{aligned}
(x + I) + ((y + I) + (z + I)) &= (x + I) + ((y + z) + I) \\
&= (x + (y + z)) + I \\
&= ((x + y) + z) + I, \text{ because } + \text{ is assocaitive in } R \\
&= ((x + y) + I) + (z + I) \\
&= ((x + I) + (y + I)) + (z + I).
\end{aligned}$$

This shows that $+$ is associative in R/I. Similarly, $+$ is commutative. Next, note that $0 + I = I$ is the additive identity and for $x + I \in R/I$, $(-x) + I$ is the additive inverse of $x + I$. As in the case of the associativity for $+$, we can show that $\cdot$ is associative.

Next, let us verify one of the distributive law. Now

$$\begin{aligned}
(x + I) \cdot ((y + I) + (z + I)) &= (x + I) \cdot ((y + z) + I) \\
&= (x(y + z)) + I \\
&= (xy + xz) + I, \text{ because distributivity holds in } R \\
&= (xy) + I) + (xz + I) \\
&= ((x + I) \cdot (y + I)) + ((x + I) \cdot (z + I)).
\end{aligned}$$

In a similar manner, we can verify the right distributive property.

Remark 8.2.21 *The quotient ring R/I can also be realised by oberving the $(I, +)$ is a normal subgroup of $(R, +)$ because the latter group is commutative. Hence, if R/I denotes the set of all cosets $x + I = \{x + a \mid a \in I\}$ for all $x \in R$, then $(R/I, +)$ is a commutative group, where*

$$(x + I) + (y + I) = (x + y) + I$$

for all $x+I$, $y+I \in R/I$. Now define multiplication on R/I by $(x+I)\cdot(y+I) = xy+I$ for all $x+I$, $y+I \in R/I$. Then $(R/I, +, \cdot)$ forms a ring.

Definition 8.2.22 *If R is a ring and I is an ideal of R, then the ring $(R/I, +, \cdot)$ is called the **quotient ring** of R by I.*

Theorem 8.2.23 *Let $n \in \mathbb{Z}$ be a fixed positive integer. Then the following conditions are equivalent.*

(i) n is prime.

(ii) $\mathbb{Z}/\langle n\rangle$ is an integral domain.

(iii) $\mathbb{Z}/\langle n\rangle$ is a field.

Proof. Note that $\langle n\rangle$ is the ideal of $\mathbb{Z}$ generated by n and $\langle n\rangle = \{nt \mid t \in \mathbb{Z}\}$.
(i) $\Rightarrow$(ii): Suppose n is prime. Let $a + \langle n\rangle, b + \langle n\rangle \in \mathbb{Z}/\langle n\rangle$. Suppose

$$(a + \langle n\rangle)(b + \langle n\rangle) = 0 + \langle n\rangle .$$

Now

$$\begin{array}{lll} & (a + \langle n\rangle)(b + \langle n\rangle) = 0 + \langle n\rangle & \\ \Rightarrow & ab + \langle n\rangle = 0 + \langle n\rangle & \\ \Rightarrow & ab \in \langle n\rangle & \\ \Rightarrow & ab = rn & \text{for some } r \in \mathbb{Z} \\ \Rightarrow & n \mid ab. & \\ \Rightarrow & \text{either } n \mid a \text{ or } n \mid b & \text{because } n \text{ is prime} \\ \Rightarrow & \text{either } a \in \langle n\rangle \text{ or } b \in \langle n\rangle & \\ \Rightarrow & \text{either } a + \langle n\rangle = 0 + \langle n\rangle \text{ or } b + \langle n\rangle = 0 + \langle n\rangle & \end{array}$$

We can now conclude that $\mathbb{Z}/\langle n\rangle$ has no zero divisors, proving that $\mathbb{Z}/\langle n\rangle$ is an integral domain.
(ii)$\Rightarrow$(iii): Because $\mathbb{Z}/\langle n\rangle$ is a finite integral domain, the result follows from Theorem 7.1.40.
(iii)$\Rightarrow$(i): Suppose n is not prime. Then $n = n_1n_2$ for some $1 < n_1 < n$ and $1 < n_2 < n$. Because $1 < n_1 < n$, we have $n_1 \notin < n >$. Similarly, $n_2 \notin < n >$. Hence $n_1 + \langle n\rangle$ and $n_2 + \langle n\rangle$ are nonzero elements of $\mathbb{Z}/\langle n\rangle$ and

$$(n_1 + \langle n\rangle)(n_2 + \langle n\rangle) = n_1n_2 + \langle n\rangle = n + \langle n\rangle = 0 + \langle n\rangle .$$

Because $\mathbb{Z}/\langle n\rangle$ is a field, $\mathbb{Z}/\langle n\rangle$ has no zero divisors. Thus, either $n_1 + \langle n\rangle = 0 + \langle n\rangle$ or $n_2 + \langle n\rangle = 0 + \langle n\rangle$, i.e., either $n_1 \in \langle n\rangle$ or $n_2 \in \langle n\rangle$,a contradiction. Therefore, n is prime. ■

We close this section by introducing the notions of nil and nilpotent ideals.

Definition 8.2.24 *Let I be an ideal of a ring R.*
*(i) I is called a **nil** ideal if each element of I is a nilpotent element.*
*(ii) I is called a **nilpotent** ideal if $I^n = \{0\}$ for some positive integer n.*

Example 8.2.25 *In the ring $\mathbb{Z}_8$, the ideal $I = \{[0], [4]\}$ is a nil ideal and also a nilpotent ideal.*

$$I^2 = \left\{ \sum_{i=1}^{k} [a_i][b_i] \mid [a_i], [b_i] \in I, k \in \mathbb{N} \right\} = \{0\}$$

because $16 \mid a_ib_i$.

From the definition, it follows that every nilpotent ideal is a nil ideal. The following example shows that the converse is not true. In this example, we construct a ring R from the rings $\mathbb{Z}_{p^n}$, $n = 1, 2, \ldots$, i.e., from the rings $\mathbb{Z}_p, \mathbb{Z}_{p^2}, \mathbb{Z}_{p^3}, \ldots$, where p is a fixed prime.

Example 8.2.26 *Let p be a fixed prime. Let R be the collection of all sequences $\{a_n\}$ such that $a_n \in \mathbb{Z}_{p^n}$ $(n \geq 1)$ and there exists a positive integer m (dependent on $\{a_n\}$) such that $a_n = [0]$ for all $n \geq m$. Define addition and multiplication on R by*

$$\begin{array}{rcl} \{a_n\} + \{b_n\} & = & \{a_n + b_n\}, \\ \{a_n\}\{b_n\} & = & \{a_nb_n\} \end{array}$$

for all $\{a_n\}, \{b_n\} \in R$. We ask the reader to verify that R is a commutative ring under these two operations, where the zero element is the sequence $\{a_n\}$ such that $a_n = [0]$ for all n and the additive inverse of the sequence $\{a_n\}$ is the sequence $\{-a_n\}$. Now in $\mathbb{Z}_{p^n}$, $[p]$ is a nilpotent element because $[p]^n = [p^n] = [0]$. Thus, for any $[r] \in \mathbb{Z}_{p^n}$, $[p][r] = [pr]$ is a nilpotent element. Therefore, we find that each element of $[p]\mathbb{Z}_{p^n}$ is a nilpotent element.
Let

$$I = \{\{[p]a_1, [p]a_2, \ldots, [p]a_n, [0], [0], \ldots\} \in R \mid n \in \mathbb{N},\ a_i \in \mathbb{Z}_{p^i}, i = 1, \ldots, n\}.$$

Then I is an ideal of R. Also, every element of I is nilpotent. Let us now show that I is not nilpotent. Suppose I is nilpotent. Then there exists a positive integer m such that $I^m = \{0\}$. Now the sequence $\{a_n\}$ such that $a_n = [p]$ for $n = 1, 2, \ldots, m+1$ and $a_n = 0$ for all $n \geq m+2$ is an element of I. Then $\{a_n\}^m = \{[0], [0], \ldots, [0], [p^m], [0], [0], \ldots\}$, where the $(m+1)$th term of this sequence is $[p^m]$ and all other terms are 0. Because $[p^m]$ is not zero in $\mathbb{Z}_{p^{m+1}}$, we find that $\{a_n\}^m \neq 0$ and $\{a_n\}^m \in I^m = \{0\}$, a contradiction. This implies that I is not nilpotent.

Theorem 8.2.27 *Let R be a commutative ring with 1 and I denote the set of all nilpotent elements of R. Then*
(i) I is a nil ideal of R,
(ii) the quotient ring R/I has no nonzero nilpotent elements.

Proof. (i) Because $0 \in I$, $I \neq \emptyset$. Let $a, b \in I$. There exist positive integers m and n such that $a^n = 0$ and $b^m = 0$. Because R is commutative, we can write

$$(a-b)^{n+m} = a^{n+m} + \cdots + (-1)^r \binom{n+m}{r} a^{n+m-r} b^r + \cdots + (-1)^{n+m} b^{n+m}.$$

The general term of the above expression is $(-1)^r \binom{n+m}{r} a^{n+m-r} b^r$, where $0 \leq r \leq m+n$. If $r \leq m$, then $n+m-r \geq n$ and hence $a^{n+m-r} = a^n a^{m-r} = 0$. Again, if $r > m$, then

$$b^r = b^{m+(r-m)} = b^m b^{r-m} = 0.$$

Therefore, we find that

$$(-1)^r \binom{n+m}{r} a^{n+m-r} b^r = 0, r = 0, 1, 2, \ldots, n+m.$$

This implies that $(a-b)^{n+m} = 0$, i.e., $a-b$ is nilpotent, so $a-b \in I$. Let $r \in R$. Then $(ra)^n = r^n a^n = r^n 0 = 0$. Because R is commutative, $(ar)^n = (ra)^n = 0$. Thus, $ar, ra \in I$. Consequently, I is an ideal of R. Because every element of I is nilpotent, I is nil.

(ii) Let $a+I$ be a nilpotent element of R/I. Then $(a+I)^n = I$ for some positive integer n. But $a^n + I = (a+I)^n$. Thus, $a^n + I = I$, which implies that $a^n \in I$. Because every element of I is nilpotent, there exists a positive integer m such that $(a^n)^m = 0$, i.e., $a^{nm} = 0$, which shows that a is nilpotent, so $a \in I$. This implies $a + I = I$. Hence, R/I has no nonzero nilpotent elements. ■

Theorem 8.2.28 *Let A and B be two nil ideals of a commutative ring R with 1. Then $A+B$ is a nil ideal.*

Proof. By Theorem 8.2.20, we know that $A+B$ is an ideal of R. Let I be the set of all nilpotent elements of R. Then $A \subseteq I$, $B \subseteq I$ and by Theorem 8.2.27, I is an ideal. Hence, $A+B \subseteq I$. Because I is nil, $A+B$ is nil. ■

Worked-Out Exercises

♢ **Exercise 1** Find all ideals of $\mathbb{Z}$.

Solution: From Worked-Out Exercise 3 (page 147), we know that the subrings of $\mathbb{Z}$ are the subsets $n\mathbb{Z}$, $n = 0, 1, 2, \ldots$. Let us now show that these subrings are precisely the ideals of $\mathbb{Z}$. If I is an ideal of $\mathbb{Z}$, then I is a subring of $\mathbb{Z}$, so $I = n\mathbb{Z}$ for some nonnegative integer n. Now, let $I = n\mathbb{Z}$ (n is a nonnegative integer). Then I is a subring. If $r \in \mathbb{Z}$, then $rI = r(n\mathbb{Z}) = n(r\mathbb{Z}) \subseteq n\mathbb{Z} = I$. Similarly, $Ir \subseteq I$. Hence, I is an ideal of $\mathbb{Z}$.

Exercise 2 Let R be a ring such that R has no zero divisors. Show that if every subring of R is an ideal of R, then R is commutative.

Solution: Let $0 \neq a \in R$. Then $C(a) = \{x \in R \mid xa = ax\}$ is a subring of R and hence an ideal of R. Thus, $ra \in C(a)$ for all $r \in R$. Let $r \in R$. Now $ara = ra^2$ implies that $(ar - ra)a = 0$. Because R has no zero divisors and $a \neq 0$, $ar - ra = 0$, so $ar = ra$. Hence, a is in the center of R. Because a is arbitrary, R is commutative.

♢ **Exercise 3** Give an example of a ring R and ideals A_i, $i \in I$, such that $A_i \cap A_j = \{0\}$ if $i \neq j$, but $A_i \cap (\sum_{j \neq i} A_j) \neq \{0\}$.

Solution: Let $R = \{0, a, b, c\}$. Define $+$ and $\cdot$ on R by

$$2a = 2b = 2c = 0,\ xy = 0,\ \text{for all } x, y \in R \text{ and}$$
$$a+b = b+a = c,\ a+c = c+a = b,\ \text{and } b+c = c+b = a.$$

Then $(R, +, \cdot)$ is a ring. Let $A_1 = \{0, a\}$, $A_2 = \{0, b\}$, and $A_3 = \{0, c\}$. Then $A_1 + A_2 = A_1 + A_3 = A_2 + A_3 = R$ and $A_1 \cap A_2 = A_1 \cap A_3 = A_2 \cap A_3 = \{0\}$.

♢ **Exercise 4** Give an example of a ring R and ideals A and B such that $AB \subset A \cap B$.

Solution: Let R be the ring of Worked-Out Exercise 3. Let $A = B = \{0, a\}$. Then $AB = \{0\} \subset \{0, a\} = A \cap B$.

♢ **Exercise 5** Characterize all commutative rings R such that R has only two ideals R and $\{0\}$.

Solution: Let R be a commutative ring such that the only ideals of R are R and $\{0\}$. Now R^2 is an ideal of R. Thus, $R^2 = \{0\}$ or $R^2 = R$.

Case 1. $R^2 = \{0\}$. Then $ab = 0$ for all $a, b \in R$. In this case, every subgroup of $(R, +)$ is an ideal. Hence, $(R, +)$ has no nontrivial subgroups, so $(R, +)$ is a cyclic group of prime order by Exercise 22 (page 94).

Case 2. $R^2 = R$. Let $0 \neq a \in R$. Then aR is an ideal of R. Hence, either $aR = \{0\}$ or $aR = R$. Suppose $aR = \{0\}$. Let $T = \langle a \rangle$. Then T is an ideal of R and $a \in T$. Thus, $T = R$. Now $aR = \{0\}$ implies that $TR = \{0\}$ and hence $R^2 = \{0\}$, which is a contradiction. Therefore, $aR = R$. Thus, for all $0 \neq a \in R$, $aR = R$. We now show that R has no zero divisors. Let a, b be two nonzero elements of R such that $ab = 0$. Let $T = \{c \in R \mid ac = 0\}$. It is easy to see that T is a nonzero ideal of R. Hence, by the hypothesis, $T = R$. This implies that $R = aR = aT = \{0\}$, a contradiction to the fact that $R = R^2 \neq \{0\}$. Consequently, R has no zero divisors. Next, for $0 \neq a \in R$, $aR = R$, so we find that $ae = a$ for some $e \in R$. Because $a \neq 0$, we must have $e \neq 0$. Also, because R has no zero divisors, $a(e^2 - e) = 0$ implies that $e^2 = e$. Now for any $b \in R$, $eb = e^2 b$ implies that $e(b - eb) = 0$ and hence $b = eb = be$. This shows that e is the identity element of R. Also, $aR = R$ implies that $e = ab$ for some $b \in R$. Hence, a^{-1} exists in R. Consequently, R is a field. So from the above two cases we conclude that either R is the zero ring with a prime number of elements or R is a field.

Exercises

1. Let $T_2(\mathbb{Z}) = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} \mid a, b, c \in \mathbb{Z} \right\}$ be the ring of all upper triangular matrices over $\mathbb{Z}$.

   (a) Prove that $I = \left\{ \begin{bmatrix} 0 & b \\ 0 & c \end{bmatrix} \mid b, c \in \mathbb{Z} \right\}$ is an ideal of $T_2(\mathbb{Z})$. Find the quotient ring $T_2(\mathbb{Z})/I$.

   (b) Prove that $I = \left\{ \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} \mid a \in \mathbb{Z} \right\}$ is an ideal of $T_2(\mathbb{Z})$. Find the quotient ring $T_2(\mathbb{Z})/I$.

2. In the ring $\mathbb{Z}_{24}$, show that $I = \{[0], [8], [16]\}$ is an ideal. Find all elements of the quotient ring $\mathbb{Z}_{24}/I$.

3. Show that the set $I = \{a + bi\sqrt{5} \mid a, b \in \mathbb{Z} \text{ and } a - b \text{ is even}\}$ is an ideal of the ring $\mathbb{Z}[i\sqrt{5}]$.

4. Let R be a ring and $a \in R$. Show that aR is a right ideal of R and Ra is a left ideal of R.

5. Let R be a ring. Let A be a left ideal of R and B be a right ideal of R. Show that AB is an ideal of R and $BA \subseteq A \cap B$.

6. Let R be a ring such that $R^2 \neq \{0\}$. Prove that R is a division ring if and only if R has no nontrivial left ideals.

7. Let R be a ring with 1. Prove that R has no nontrivial left ideals if and only if R has no nontrivial right ideals.

8. Let I_1, I_2 be ideals of a ring R. Prove that $I_1 \cup I_2$ is an ideal of R if and only if either $I_1 \subseteq I_2$ or $I_2 \subseteq I_1$.

9. Let I and J be ideals of a ring R. Prove that $I + J$ is an ideal of R and that $I + J = \langle I \cup J \rangle$, the ideal of R generated by $I \cup J$.

10. Let I be an ideal of a commutative ring R and $a \in R$. Prove that $\langle I \cup \{a\} \rangle = \{i + ra + na \mid i \in I, r \in R, n \in \mathbb{Z}\}$.

11. Let m and n be positive integers in $\mathbb{Z}$. Prove that

    (a) $\langle m, n \rangle = \langle m \rangle + \langle n \rangle = \langle d \rangle$, where d is the greatest common divisor of m and n;

    (b) $\langle m \rangle \cap \langle n \rangle = \langle q \rangle$, where q is the least common multiple of m and n.

12. Find all ideals of the Cartesian product $F_1 \times F_2$ of two fields F_1 and F_2.

13. Consider the Cartesian product ring $R_1 \times R_2$ of the rings R_1 and R_2.

    (a) If I_1 is an ideal of R_1 and I_2 is an ideal of R_2, prove that $I_1 \times I_2$ is an ideal of $R_1 \times R_2$.

    (b) Suppose R_1 and R_2 are with 1 and I is an ideal of $R_1 \times R_2$. Does there exist ideals I_1 of R_1 and I_2 of R_2 such that $I = I_1 \times I_2$?

14. Let R be an ideal of a ring R. Prove that the quotient ring R/I is a commutative ring if and only if $ab - ba \in I$ for all $a, b \in R$.

15. Let $T = \{\frac{a}{b} \mid \frac{a}{b} \in \mathbb{Q}$, a and b are relatively prime and 5 does not divide $b\}$. Show that T is a ring under the usual addition and multiplication. Also, prove that $I = \{\frac{a}{b} \in T \mid 5$ divides $a\}$ is an ideal of T and the quotient ring T/I is a field.
16. Let I be an ideal of a ring R. Prove that if R is a commutative ring with identity, then R/I is a commutative ring with identity. If R has no zero divisors, is the same necessarily true for R/I?
17. Let I be an ideal of a commutative ring R. Define the **annihilator** of I to be the set

$$\text{ann}I = \{r \in R \mid ra = 0 \text{ for all } a \in I\}.$$

Prove that ann$I$ is an ideal of R.
18. In the ring $\mathbb{Z}_{20}$, prove that $I = \{[n] \mid n$ is even$\}$ is an ideal. Find ann$I$.
19. In the ring $\mathbb{Z}[i]$, show that $I = \{a + bi \mid a, b \in \mathbb{Z}$ and a, b are even$\}$ is an ideal. Find ann$I$.
20. In a commutative regular ring R with 1, prove that every principal ideal I is generated by an idempotent and for every principal ideal I, there exists a principal ideal J such that $R = I + J$ and $I \cap J = \{0\}$.
21. Prove that every ideal of a regular ring is regular.
22. Prove that a ring R is regular if and only if every principal left ideal of R is generated by an idempotent.
23. Prove that in a commutative regular ring with 1 every finitely generated ideal is a principal ideal.
24. In a ring R, prove that $\{0\}$ is the only nilpotent ideal if and only if for all ideals A and B of R, $AB = \{0\}$ implies $A \cap B = \{0\}$.
25. Let R be a ring and $f : R \to [0, 1]$ be such that

$$\begin{aligned} f(a-b) &\geq \min\{f(a), f(b)\}, \\ f(rb) &\geq f(b) \end{aligned}$$

for all $a, b, r \in R$. Prove the following:
(i) $f(0) \geq f(a)$ for all $a \in R$;
(ii) $f(a) = f(-a)$ for all $a \in R$;
(iii) for all $t \in \mathcal{I}(f)$, $R_t = \{x \in R \mid f(x) \geq t\}$ is a left ideal of R;
(iv) $R_0 = \{a \in R \mid f(a) = f(0)\}$ is a left ideal of R.
26. Let R be a ring. A relation ρ on R is called a congruence relation on the ring R if ρ is an equivalence relation on R and for all $a, b, c \in R$, $a\rho b$ implies that $ac\rho bc$, $ca\rho cb$, and $(a+c)\rho(b+c)$. Let I be an ideal of R and ρ be the relation on R defined by $a\rho b$ if and only if $a - b \in I$. Show that ρ is a congruence relation on R.
27. In each of the following exercises, write the proof if the statement is true; otherwise, give a counterexample.
(i) If $\{I_i \mid i \in \mathbb{N}\}$ is a collection of ideals of R, then $\cup_{i\in\mathbb{N}} I_i$ is an ideal of R.
(ii) $\mathbb{Z}$ is a subring of $\mathbb{R}$, but not an ideal of $\mathbb{R}$.
(iii) If I is a nontrivial ideal of an integral domain R, then the quotient ring R/I is an integral domain.

8.3 Homomorphisms and Isomorphisms

In this section, we introduce the ideas of homomorphisms and isomorphisms of rings. These concepts are the analogs of homomorphisms and isomorphisms for groups.

Definition 8.3.1 *Let $(R, +, \cdot)$ and $(R', +', \cdot')$ be rings and f a function from R into R'. Then f is called a **homomorphism** of R into R' if*

$$\begin{aligned} f(a+b) &= f(a) +' f(b), \\ f(a \cdot b) &= f(a) \cdot' f(b) \end{aligned}$$

for all $a, b \in R$.

A homomorphism f of a ring R into a ring R' is called
(i) a **monomorphism** if f is one-one,
(ii) an **epimorphism** if f is onto R', and
(iii) an **isomorphism** if f is one-one and maps R onto R'.

If f is an isomorphism of a ring R onto a ring R', then f^{-1} is an isomorphism of R' onto R.
An isomorphism of a ring R onto R is called an **automorphism**.

Definition 8.3.2 *Two rings R and R' are said to be* ***isomorphic*** *if there exists an isomorphism of R onto R'.*

We write $R \simeq R'$ when R and R' are isomorphic.

When speaking of two rings R and R', from now on we usually use the operations $+$ and $\cdot$ for both rings. Let $f : R \to R'$ be a homomorphism of rings. Because f preserves $+$, f is a also a homomorphism of the groups $(R, +)$ and $(R', +)$. Hence, we can immediately apply Theorem 5.1.2 to conclude that f maps 0 to $0'$, i.e., $f(0) = 0'$, and for all $a \in R$, $-f(a) = f(-a)$. We list some properties of homomorphisms in the following theorem. The proofs are similar to the proof of Theorem 5.1.2, so we leave them as an exercise for the reader.

Theorem 8.3.3 *Let f be a homomorphism of a ring R into a ring R'. Then the following assertions hold.*
(i) $f(0) = 0'$, where $0'$ is the zero of R'.
(ii) $f(-a) = -f(a)$ for all $a \in R$.
(iii) $f(R) = \{f(a) \mid a \in R\}$ is a subring of R'.
(iv) If R is commutative, then $f(R)$ is commutative.
Suppose R has an identity and $f(R) = R'$, i.e., f is onto R'. Then
(v) R' has an identity, namely, $f(1)$.
(vi) If $a \in R$ is a unit, then $f(a)$ is a unit in R' and

$$f(a)^{-1} = f(a^{-1}). \blacksquare$$

We point out that in (v) of Theorem 8.3.3, if f is not onto, then R' may or may not have an identity. Even if R' has an identity, the identity of R need not map onto the identity of R'. We illustrate this point later in Example 8.3.7.

Definition 8.3.4 *Let f be a homomorphism of a ring R into a ring R'. Then the* ***kernel*** *of f, written Ker f, is defined to be the set*

$$Ker\ f = \{a \in R \mid f(a) = 0'\},$$

where $0'$ denote the additive identity of R'.

From Theorem 8.3.3, we know that $0 \in$ Ker f. Hence, Ker $f \neq \emptyset$.

Example 8.3.5 *(i) The identity map of a ring R is a homomorphism (in fact, an isomorphism). Its kernel is $\{0\}$.*
(ii) Let R and R' be rings and $f : R \to R'$ be defined by $f(a) = 0'$ for all $a \in R$. Then f is a homomorphism of R into R' and Ker $f = R$.

Example 8.3.6 *Let f be the mapping from $\mathbb{Z}$ onto $\mathbb{Z}_n$ defined by $f(a) = [a]$ for all $a \in \mathbb{Z}$. From Example 5.1.4, $f(a + b) = f(a) +_n f(b)$ for all $a, b \in \mathbb{Z}$. Also, $f(a \cdot b) = [ab] = [a] \cdot_n [b] = f(a) \cdot_n f(b)$ for all $a, b \in \mathbb{Z}$. Thus, f is a homomorphism of $\mathbb{Z}$ onto $\mathbb{Z}_n$. As in Example 5.1.4, Ker $f = \{qn \mid q \in \mathbb{Z}\}$.*

In the following example, we show that if f is a homomorphism from a ring R with 1 into a ring R' with 1 and f is not onto, then the identity of R need not map onto the identity of R'.

Example 8.3.7 *Consider the direct sum $\mathbb{Z} \oplus \mathbb{Z}$ of $\mathbb{Z}$ with itself (see Exercise 17, page 140). Define $f : \mathbb{Z} \to \mathbb{Z} \oplus \mathbb{Z}$ by $f(a) = (a, 0)$ for all $a \in \mathbb{Z}$. From the definition of f, f is well defined. Now for all $a, b \in \mathbb{Z}$,*

$$f(a + b) = (a + b, 0) = (a, 0) + (b, 0) = f(a) + f(b)$$

and

$$f(ab) = (ab, 0) = (a, 0)(b, 0) = f(a)(b).$$

Thus, f is a homomorphism. Also, Ker $f = \{0\}$. Now $f(1) = (1, 0)$, but $(1, 1)$ is the identity of $\mathbb{Z} \oplus \mathbb{Z}$. Therefore, the identity of $\mathbb{Z}$ does not map onto the identity of $\mathbb{Z} \oplus \mathbb{Z}$.

Consider the rings $\mathbb{Z}$ and $\mathbb{Q}$. Suppose $\mathbb{Z} \simeq \mathbb{Q}$. Then the groups $(\mathbb{Z}, +)$ and $(\mathbb{Q}, +)$ are isomorphic. However, this is not possible because $(\mathbb{Z},+)$ is a cyclic group and $(\mathbb{Q},+)$ is not a cyclic group. In the following example, we give another argument to show that $\mathbb{Z}$ is not isomorphic to $\mathbb{Q}$.

Example 8.3.8 *Suppose $\mathbb{Z} \simeq \mathbb{Q}$. Let $f : \mathbb{Z} \to \mathbb{Q}$ be an isomorphism. Then $f(1) = 1$ and $f(0) = 0$. Let n be a positive integer. Then*

$$f(n) = f(\underbrace{1 + \cdots + 1}_{n \text{ times}}) = f(1) + f(1) + \cdots + f(1) = nf(1) = n1 = n.$$

Now suppose that n is a negative integer. Let $n = -m$, where m is positive. Then $f(n) = f(-m) = f(-1 - 1 - \cdots - 1) = -f(1) - f(1) - \cdots - f(1) = m(-f(1)) = -mf(1) = -m1 = -m = n$. Hence, $f(n) = n$ for all $n \in \mathbb{Z}$. Let $0 \neq \frac{a}{b} \in \mathbb{Q} \backslash \mathbb{Z}$. Because f is onto $\mathbb{Q}$, there exists $n \in \mathbb{Z}$ such that $\frac{a}{b} = f(n) = n$, which is a contradiction. Hence, $\mathbb{Q}$ is not isomorphic to $\mathbb{Z}$.

In the following example, we consider two rings which look similar, but which are not isomorphic.

Example 8.3.9 *In this example, we show that the ring $\mathbb{Z}[\sqrt{3}] = \{a + b\sqrt{3} \mid a, b \in \mathbb{Z}\}$ and the ring $\mathbb{Z}[\sqrt{5}] = \{a + b\sqrt{5} \mid a, b \in \mathbb{Z}\}$ are not isomorphic.*

Suppose there exists an isomorphism $f : \mathbb{Z}[\sqrt{3}] \to \mathbb{Z}[\sqrt{5}]$. Now $3 = (0 + \sqrt{3})^2$. Thus,

$$f(3) = f((\sqrt{3})^2) = (f(\sqrt{3}))^2.$$

Because f is an isomorphism, we have $f(1) = 1$. This implies that $f(3) = 3$. Hence,

$$3 = f(3) = (f(\sqrt{3}))^2.$$

Because $f(\sqrt{3}) \in \mathbb{Z}[\sqrt{5}]$, $f(\sqrt{3}) = a + b\sqrt{5}$ for some $a + b\sqrt{5} \in \mathbb{Z}[\sqrt{5}]$. Therefore,

$$3 = (a + b\sqrt{5})^2.$$

This implies that

$$3 = a^2 + 5b^2 + 2ab\sqrt{5}.$$

If $ab = 0$, then $3 = a^2 + 5b^2$. But there do not exist integers a and b such that $ab = 0$ and $3 = a^2 + 5b^2$.

If $ab \neq 0$, then $\sqrt{5} = \frac{3 - a^2 - 5b^2}{2ab} \in \mathbb{Q}$, which is a contradiction. Hence, $\mathbb{Z}[\sqrt{3}]$ and $\mathbb{Z}[\sqrt{5}]$ are not isomorphic.

The next example shows that the ring $\mathbb{Z}_n$ and the ring $\mathbb{Z}/\langle n \rangle$ are isomorphic.

Example 8.3.10 *Consider the ideal $\langle n \rangle$ generated by a fixed positive integer $n \in \mathbb{Z}$. By Corollary 8.2.11, $\langle n \rangle = \{qn \mid q \in \mathbb{Z}\}$. The cosets of $\langle n \rangle$ in $\mathbb{Z}$ are $a + \langle n \rangle = \{a + qn \mid q \in \mathbb{Z}\}$, where $a \in \mathbb{Z}$. Now*

$$\mathbb{Z}/\langle n \rangle = \{a + \langle n \rangle \mid \ a \in \mathbb{Z}\}.$$

Define $f : \mathbb{Z}_n \to \mathbb{Z}/\langle n \rangle$ by $f([a]) = a + \langle n \rangle$ for all $[a] \in \mathbb{Z}_n$. We recall that f is an isomorphism of $(\mathbb{Z}_n, +_n)$ onto $(\mathbb{Z}/\langle n \rangle, +)$ (Example 5.1.15). Now

$$f([a] \cdot_n [b]) = f([ab]) = ab + \langle n \rangle = (a + \langle n \rangle)(b + \langle n \rangle) = f([a])f([b]).$$

Thus, f is a ring isomorphism of $\mathbb{Z}_n$ onto $\mathbb{Z}/\langle n \rangle$.

Theorem 8.3.11 *Let f be a homomorphism of a ring R into a ring R'. Then Ker f is an ideal of R.*

Proof. Because $0 \in$ Ker f, Ker $f \neq \emptyset$. Let $a, b \in$ Ker f. Then $f(a - b) = f(a) - f(b) = 0' - 0' = 0'$, so $a - b \in$ Ker f. Let $r \in R$. Then $f(ra) = f(r) \cdot f(a) = f(r) \cdot 0' = 0'$, so $ra \in R$. Similarly, $ar \in$ Ker f. Hence, Ker f is an ideal of R.
■

In the remainder of the section, we consider isomorphism theorems which are parallel to those for groups (Section 5.2).

Theorem 8.3.12 *Let R be a ring and I be an ideal of R. Define the mapping $g : R \to R/I$ by $g(a) = a + I$ for all $a \in R$. Then g is a homomorphism, called the* ***natural homomorphism****, of R onto R/I. Furthermore, Ker $g = I$.*

Proof. Now for all $a, b \in R$,

$$g(a+b) = (a+b) + I = (a+I) + (b+I) = g(a) + g(b)$$

and

$$g(ab) = ab + I = (a+I)(b+I) = g(a)g(b).$$

That Ker $g = I$ follows from Theorem 5.1.12 in group theory. ■

Theorem 8.3.13 *Let f be a homomorphism of a ring R onto a ring R' and I be an ideal of R contained in Ker f. Let g be the natural homomorphism of R onto R/I. Then there exists a unique homomorphism h of R/I onto R' such that $f = h \circ g$. Furthermore, h is one-one if and only if $I =$ Ker f.*

Proof. Once again, we use the work already done for groups. Define $h : R/I \to R'$ by $h(a+I) = f(a)$ for all $a \in R$. We have the desired results by Theorem 5.2.1, once we verify that h preserves multiplication. Now

$$h((a+I)(b+I)) = h(ab+I) = f(ab) = f(a)f(b) = h(a+I)h(b+I).$$

■

The proof of the following theorem is similar to that of the first isomorphism theorem for groups. We omit the proof. This theorem is also known as **the fundamental theorem of homomorphisms** for rings.

Theorem 8.3.14 (First Isomorphism Theorem) *Let f be a homomorphism of a ring R into a ring R'. Then $f(R)$ is an ideal of R' and*

$$R/Ker\ f \simeq f(R). \blacksquare$$

We state the following theorem without proof. Its proof is a direct translation of the proof of the corresponding theorem for groups.

Theorem 8.3.15 (Correspondence Theorem) *Let f be a homomorphism of a ring R onto a ring R'. Then f induces a one-one inclusion preserving correspondence between the ideals of R containing Ker f and the ideals of R' in such a way that if I is an ideal of R containing Ker f, then $f(I)$ is the corresponding ideal of R', and if I' is an ideal of R', then $f^{-1}(I')$ is the corresponding ideal of R.* ■

An example similar to Example 5.2.13 can be developed to illustrate Theorem 8.3.15

The next two isomorphism theorems for rings correspond to Theorems 5.2.8 and 5.2.6, respectively.

Theorem 8.3.16 *Let f be a homomorphism of a ring R onto a ring R', I be an ideal of R such that $I \supseteq$ Ker f, g, and g' be the natural homomorphisms of R onto R/I and R' onto $R'/f(I)$, respectively. Then there exists a unique isomorphism h of R/I onto $R'/f(I)$ such that $g' \circ f = h \circ g$.* ■

Corollary 8.3.17 *Let I_1, I_2 be ideals of a ring R such that $I_1 \subseteq I_2$. Then*

$$(R/I_1)/(I_2/I_1) \simeq R/I_2. \blacksquare$$

Theorem 8.3.18 *If I and J are ideals of the ring R, then $I/(I \cap J) \simeq (I+J)/J$.* ■

Worked-Out Exercises

♢ **Exercise 1** Show that the function $f : \mathbb{Z}_6 \to \mathbb{Z}_{10}$ defined by $f([a]) = 5[a]$ for all $[a] \in \mathbb{Z}_6$ is a ring homomorphism of $\mathbb{Z}_6$ into $\mathbb{Z}_{10}$.

Solution: We first show that f is well defined. Let $[a] = [b]$ in $\mathbb{Z}_6$. Then $a - b$ is divisible by 6. Thus, $a = 6k + b$ for some $k \in \mathbb{Z}$. Now $5a = 30k + 5b$ shows that $5[a] = [5a] = [30k + 5b] = [30k] +_{10} [5b] = [0] +_{10} 5[b] = 5[b]$ in $\mathbb{Z}_{10}$. Therefore, $f([a]) = f([b])$. Thus, we find that f is well defined. Let $[a], [b] \in \mathbb{Z}_6$. Then $f([a] +_6 [b]) = f([a+b]) = 5[a+b] = 5([a] +_{10} [b]) = 5[a] +_{10} 5[b] = f(a) +_{10} f(b)$ and $f([a] \cdot_6 [b]) = f([ab]) = 5[ab] = 25[ab]$ (because $\mathbb{Z}_{10}$ is of characteristic 10) $= (5[a]) \cdot_{10} (5[b]) = f(a) \cdot_{10} f(b)$. Hence, f is a homomorphism.

♢ **Exercise 2** Let $\mathbb{R}$ be the field of real numbers. Let α be an automorphism of $\mathbb{R}$. Show that $\alpha(x) = x$ for all $x \in \mathbb{R}$.

Solution: Because α is an automorphism of $\mathbb{R}$, $\alpha(0) = 0$, and $\alpha(1) = 1$. Let $n \in \mathbb{N}$. Then $\alpha(n) = \alpha(1+1+\cdots+1) = \alpha(1) + \alpha(1) + \cdots + \alpha(1) = 1 + 1 + \cdots + 1 = n$. Now let $m \in \mathbb{Z}$ and $m < 0$. Let $n = -m > 0$. Then $\alpha(m) = \alpha(-n) = -\alpha(n) = -n = m$. This shows that $\alpha(x) = x$ for all $x \in \mathbb{Z}$. Let $\frac{p}{q} \in \mathbb{Q}$. Then $\alpha(\frac{p}{q}) = \alpha(pq^{-1}) = \alpha(p)\alpha(q^{-1}) = p\alpha(q)^{-1} = pq^{-1} = \frac{p}{q}$. This shows that $\alpha(x) = x$ for all $x \in \mathbb{Q}$. Let $x \in \mathbb{R}$ be such that $x \geq 0$. Then $x = y^2$ for some $y \in \mathbb{R}$. Thus, $\alpha(x) = \alpha(y^2) = \alpha(yy) = \alpha(y)\alpha(y) = \alpha(y)^2 \geq 0$. Now let $a, b \in \mathbb{R}$ be such that $a \geq b$. Then $a - b \geq 0$. Hence, $\alpha(a-b) \geq 0$, so $\alpha(a) - \alpha(b) \geq 0$, i.e., $\alpha(a) \geq \alpha(b)$. Therefore, α is order preserving. We now show that α is continuous. Let $\epsilon \in \mathbb{R}$ and $\epsilon > 0$. Because α is onto $\mathbb{R}$, there exists $\delta > 0$ such that $\alpha(\delta) = \epsilon$. Now let $x, y \in \mathbb{R}$ be such that $|x - y| < \delta$. Thus,

$$-\delta < x - y < \delta.$$

Because α is order preserving,

$$\alpha(-\delta) < \alpha(x - y) < \alpha(\delta).$$

Therefore,

$$-\epsilon < \alpha(x - y) < \epsilon,$$

so

$$-\epsilon < \alpha(x) - \alpha(y) < \epsilon.$$

This implies that

$$|\alpha(x) - \alpha(y)| < \epsilon.$$

Hence, α is continuous. Now let $x \in \mathbb{R}$. Because $\mathbb{Q}$ is dense in $\mathbb{R}$, there exists a sequence $\{a_n\}$ of rational numbers such that

$$\lim_{n\to\infty} a_n = x.$$

Because α is continuous,

$$\alpha(x) = \alpha(\lim_{n\to\infty} a_n) = \lim_{n\to\infty} \alpha(a_n) = \lim_{n\to\infty} a_n = x,$$

proving the result.

♢ **Exercise 3** Let R be a ring with 1. If the characteristic of R is 0, then show that R contains a subring isomorphic to $\mathbb{Z}$.

Solution: Let $T = \{n1 \mid n \in \mathbb{Z}\}$. Because $0 = 01 \in T$, $T \neq \emptyset$. Let $a = n1$ and $b = m1$ be two elements of T. Then $a - b = n1 - m1 = (n-m)1$ and $ab = (n1)(m1) = (nm)1$. Hence, $a - b, ab \in T$. Thus, T is a subring of R. Suppose n, m are two integers such that $n1 = m1$. If $n > m$, then $(n-m)1 = 0$. This contradicts the assumption that R is of characteristic 0. Similarly, $m > n$ also leads to a contradiction. Hence, $n = m$. Thus, we find that for each $a \in T$, there exists a unique integer n such that $a = n1$. Hence, the mapping $f : \mathbb{Z} \to T$ defined by $f(n) = n1$ is an isomorphism.

Exercise 4 Let p be a prime integer. Show that there are only two nonisomorphic rings of p elements.

Solution: It is known that $(\mathbb{Z}_p, +_p)$ is the only group of order p (up to isomorphism). Define $\odot_1$ and $\odot_2$ on $\mathbb{Z}_p$ by $[a] \odot_1 [b] = [0]$ and $[a] \odot_2 [b] = [ab]$ for all $[a], [b] \in \mathbb{Z}_p$. Now $\odot_1$ and $\odot_2$ are well defined and $(\mathbb{Z}_p, +_p, \odot_1)$ and $(\mathbb{Z}_p, +_p, \odot_2)$ are rings. Let R be a ring with p elements. Then $(R, +) \simeq (\mathbb{Z}_p, +_p)$. If $R \not\simeq (\mathbb{Z}_p, +_p, \odot_1)$, then the multiplication of R is not $\odot_1$. Let $[a]$ be a generator of $(\mathbb{Z}_p, +_p)$. Now $[a]^2 = n[a]$ for some nonzero integer n. There exists an integer m such that $mn \equiv_p 1$. Let $[b] = m[a]$. Then $[b]^2 = m^2[a]^2 = m^2 n[a] = m[a] = [b]$. Let g be an isomorphism from $(\mathbb{Z}_p, +_p)$ onto $(R, +)$. Define $f : \mathbb{Z}_p \to R$ by $f([u]) = ug([b])$ for all $[u] \in \mathbb{Z}_p$.Then $f([u] +_p [v]) = f([u+v]) = (u+v)g([b]) = ug([b]) + vg([b]) = f([u]) + f([v])$ and $f([u] \odot_2 [v]) = f([uv]) = (uv)g([b]) = uvg([b]^2) = uvg([b])g([b]) = ug([b])vg([b]) = f([u])f([v])$. Hence, f is a ring homomorphism. Let $c \in R$. Then there exists $[u] \in \mathbb{Z}_p$ such that $g([u]) = c$. Now $[u] = t[a]$ for some $t \in \mathbb{Z}$. Thus, $f([tn]) = tng([b]) = tn\ g(m[a]) = tg(mn[a]) = tg([a]) = g(t[a]) = g([u]) = c$. Hence, f is onto R. Because $|\mathbb{Z}_p| = |R|$, it follows that f is one-one. Thus, f is an isomorphism.

Exercises

1. Let R denote the set of all 2×2 matrices of the form $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$, where a and b are real numbers. Prove that R is a ring and the function $a + bi \to \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$ is an isomorphism of $\mathbb{C}$ onto R.
2. Define the binary operations $\oplus$ and $\odot$ on $\mathbb{Z}$ by $a \oplus b = a + b - 1$ and $a \odot b = a + b - ab$ for all $a, b \in \mathbb{Z}$. Show that $(\mathbb{Z}, \oplus, \odot)$ is a ring isomorphic to the ring $(\mathbb{Z}, +, \cdot)$.

3. (i) Show that the rings $\mathbb{R}$ and $\mathbb{Q}$ are not isomorphic.
   (ii) Show that the rings $\mathbb{R}$ and $\mathbb{C}$ are not isomorphic.
   (iii) Are the rings $\mathbb{Z}_6$ and $\mathbb{Z}_3 \times \mathbb{Z}_2$ isomorphic?
4. Let $T_2(\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mid a, b, c \in \mathbb{Z} \right\}$ be the ring of all upper triangular matrices over $\mathbb{Z}$. Define $f : T_2(\mathbb{Z}) \to \mathbb{Z}$ by for all $\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \in T_2(\mathbb{Z})$,
$$f(\begin{pmatrix} a & b \\ 0 & c \end{pmatrix}) = a.$$
   (i) Show that f is a homomorphism.
   (ii) Is f an epimorphism?
   (iii) Is f an isomorphism?
   (iv) Find Ker f.
5. Does there exist an epimorphism from the ring $\mathbb{Z}_{24}$ onto the ring $\mathbb{Z}_7$?
6. Show that there does not exist a monomorphism from the ring $\mathbb{Z}_6$ into the ring $\mathbb{Z}_{11}$.
7. Show that the ring $2\mathbb{Z}$ is not isomorphic to the ring $3\mathbb{Z}$.
8. Let R be a Boolean ring. If $\{0\}$ and R are the only ideals of R, prove that $R \simeq \mathbb{Z}_2$.
9. Show that the ring $\mathbb{Z}$ is not isomorphic to any proper subring of $\mathbb{Z}$.
10. Is the ring $\mathbb{Q}[\sqrt{2}]$ isomorphic to the ring $\mathbb{Q}[\sqrt{3}]$?
11. Let $f : R \to S$ be a nontrivial homomorphism from a field R onto a ring S. Prove that S is a field.
12. Let R be a ring with 1. If R is of characteristic $n > 0$, show that R contains a subring isomorphic to the ring $\mathbb{Z}_n$.
13. Show that there exist only two homomorphisms from $\mathbb{R}$ into $\mathbb{R}$.
14. Prove that every ring R is isomorphic to a subring of $M_n(R)$, the ring of $n \times n$ matrices over R.
15. Let f be a homomorphism of a ring R onto a ring R'. Prove that
   (i) if I is an ideal of R, then $f(I)$ is an ideal of R';
   (ii) if I' is an ideal of R', then $f^{-1}(I')$ is an ideal of R and $f^{-1}(I') \supseteq$ Ker f;
   (iii) if R is commutative and I and J are two ideals of R, then $f(I+J) = f(I)+f(J)$ and $f(IJ) = f(I)f(J)$.
16. In each of the following exercises, write the proof if the statement is true; otherwise, give a counterexample.
   (i) There exist only two homomorphisms from the ring of integers into itself.
   (ii) The mapping $f : \mathbb{Z} \to \mathbb{Z}$ defined by $f(n) = 3n$ is a group homomorphism, but not a ring homomorphism.
   (iii) The only isomorphism of a ring R onto itself is the identity mapping of R.
   (iv) Let R be a ring with 1. Let $f : R \to S$ be a ring homomorphism. Then $f(1)$ is the identity element of S.
   (v) A nonzero homomorphism from a field into a ring with more than one element is a monomorphism.
   (vi) Every nontrivial homomorphic image of an integral domain is an integral domain.

Chapter 9

Ring Embeddings

9.1 Embedding of Rings

Sometimes it is worthwhile to study the properties of a ring by considering it as a subring of some ring with more ring properties than itself. A ring without identity lacks important arithmetic properties, in particular, a fundamental theorem of arithmetic. As another example, in the ring $\mathbb{E}$ of even integers, we cannot say that 2 divides 2 because $1 \notin \mathbb{E}$. Now $\mathbb{E}$ is a subring of $\mathbb{Z}$ and $1 \in \mathbb{Z}$. In $\mathbb{Z}$, it is true that 2 divides 2. The main aim of this section is to embed a ring into a suitable ring with additional properties. The main feature of this section is that any integral domain can be embedded in a field. The proof of this result yields a rigorous construction of the rational numbers from the integers.

Definition 9.1.1 *A ring R is said to be **embedded** in a ring S if there exists a monomorphism of R into S.*

From the above definition, it follows that a ring R can be embedded in a ring S if there exists a subring T of S such that $R \simeq T$.

In the next theorem, we show that any ring R can be embedded in a ring with identity.

Theorem 9.1.2 *Any ring R can be embedded in a ring S with 1 such that R is an ideal of S. If R is commutative, then S is commutative.*

Proof. Set $S = R \times \mathbb{Z}$. Define addition and multiplication as follows:

$$\begin{aligned}(a,m)+(b,n) &= (a+b, m+n),\\ (a,m)\cdot(b,n) &= (ab+na+mb, mn)\end{aligned}$$

for all $a, b \in R$ and $m, n \in \mathbb{Z}$. (Here na means a adds to itself n times if n is positive, $-a$ adds to itself $|n|$ times if n is negative, and $0a = 0$.) Then S forms a ring under these definitions of addition and multiplication, a fact we ask the reader to prove in the exercises. We do note that $(0,0)$ is the additive identity and that $(0,1)$ is the multiplicative identity of S.

Consider the subset $R \times \{0\}$ of S. Because $(0,0) \in R \times \{0\}$, $R \times \{0\} \neq \emptyset$. Also, for all $(a,0), (b,0) \in R \times \{0\}$,

$$(a,0) - (b,0) = (a-b, 0) \in R \times \{0\},$$

and

$$(a,0) \cdot (b,0) = (ab, 0) \in R \times \{0\}.$$

Thus, $R \times \{0\}$ is a subring of S. Now for all $(a,0) \in R \times \{0\}$ and $(c,n) \in S$,

$$(a,0) \cdot (c,n) = (ac + na, 0) \in R \times \{0\}$$

and

$$(c,n) \cdot (a,0) = (ca + na, 0) \in R \times \{0\}.$$

This proves that $R \times \{0\}$ is an ideal of S.

Now define $f : R \to R \times \{0\}$ by $f(a) = (a,0)$ for all $a \in R$. Then f is an isomorphism of R onto $R \times \{0\}$, so $R \simeq R \times \{0\}$. Therefore, R can be embedded in S. By identifying $a \in R$ with $(a,0) \in R \times \{0\}$, we can regard

R to be an ideal of S. To show that the commutativity of R implies that of S, let $(a,m),(b,n) \in S$ and R be commutative. Then

$$\begin{aligned}(a,m)\cdot(b,n) &= (ab+na+mb,mn)\\ &= (ba+mb+na,nm) \quad (\text{because } R \text{ iscommutative}, ab=ba)\\ &= (b,n)\cdot(a,m).\end{aligned}$$

Hence, S is commutative. ■

Our main objective in this section is to embed a ring in a field. By Theorem 9.1.2, every ring can be embedded in a ring with identity. If S were a field, then S is commutative and has no zero divisors. This in turn implies that R is commutative and has no zero divisors. Thus, if we were to embed a ring R in a field S, then R must have at least these two properties, i.e., R must be commutative and have no zero divisors. In the next theorem, we embed a commutative ring with no zero divisors into an integral domain and then we will embed an integral domain in a field.

Theorem 9.1.3 *Let R be a commutative ring with no zero divisors. Then R can be embedded in an integral domain.*

Proof. Let S be the ring as defined in Theorem 9.1.2. Let A be the annihilator of R in S. Then A is an ideal of S by Exercise 17 (page 158). If $R \cap A = \{0\}$, then the natural homomorphism of R onto the quotient ring S/A must map R one-one into S/A, i.e., R can be embedded in S/A. We now show that $R \cap A = \{0\}$ and that S/A is an integral domain. Let $a \in R \cap A$. Then $ar = 0$ for all $r \in R$. Because R has no zero divisors, $a = 0$. Therefore, $R \cap A = \{0\}$. Let $b + A$, $c + A \in S/A$. If $(b + A)(\ c + A) = 0 + A$, then $bc \in A$. Thus, $(bc)r = 0$ for all $r \in R$. Suppose $c + A \neq 0 + A$, i.e., $c \notin A$. Then there exists $r \in R$ such that $cr \neq 0$. Because R is an ideal of S, $cr \in R$, and for all $s \in R$, $bs \in R$. Now

$$(cr)(bs) = (bcr)s = 0s = 0.$$

Also, R has no zero divisors and $cr \neq 0$. Therefore, we must have $bs = 0$. This implies that $b \in A$, so $b+A = 0+A$. Hence, S/A is an integral domain. ■

Suppose we are given the ring of integers $\mathbb{Z}$ and we are asked to construct the rational numbers from $\mathbb{Z}$. We can think of any integer as $n/1$, i.e., n divided by 1. However, we must somehow pick up the fractions which cannot be reduced to having a 1 for a denominator. One idea that suggests itself is to consider the Cartesian product $\mathbb{Z} \times \mathbb{Z}$ and consider the first component of the elements of $\mathbb{Z} \times \mathbb{Z}$ as the numerator and the second component as the denominator. However, the ordered pairs $(3, 2)$ and $(6, 4)$ are distinct. A common technique used in mathematics suggests putting these elements in the same equivalence class so that they become "equal." This is precisely what we shall do. Let's also remember not to have 0 in the denominator.

Theorem 9.1.4 *Any integral domain R can be embedded in a field.*

Proof. Let $S = R \times (R \backslash \{0\})$. Define the relation $\sim$ on S by for all $(a,b),(c,d) \in S$, $(a,b) \sim (c,d)$ if and only if $ad = bc$. Then $\sim$ is an equivalence relation. The reflexive and symmetric properties are immediate. Suppose that $(a, b) \sim (c,\ d)$ and $(c, d) \sim (e,\ f)$. Then $ad = bc$ and $cf = de$. This implies that $adf = bcf$ and $bcf = bde$, so $adf = bde$. Canceling d, we obtain $af = be$, i.e., $(a, b) \sim (e,\ f)$. Hence, $\sim$ is transitive. Now $\sim$ partitions S into equivalence classes. Denote the equivalence class $\{(c, d) \in S \mid (c, d) \sim (a, b)\}$ by a/b. Set

$$F = \{a/b \mid (a,b) \in S\}.$$

Define $+$ and $\cdot$ on F as follows:

$$\begin{aligned}a/b + c/d &= (ad+bc)/bd,\\ a/b \cdot c/d &= ac/bd\end{aligned}$$

for all $a/b, c/d \in F$. We show that $+$ is well defined. Let $a/b, c/d, a'/b', c'/d' \in F$. Suppose $a/b = a'/b'$ and $c/d = c'/d'$. Then $ab' = ba'$ and $cd' = dc'$. Therefore, $ab'dd' = ba'dd'$ and $cd'bb' = dc'bb'$. Hence,

$$ab'dd' + cd'bb' = ba'dd' + dc'bb'.$$

This implies that

$$(ad+bc)b'd' = bd(a'd' + b'c').$$

Thus,

$$(ad+bc, bd) \sim (a'd' + b'c', b'd').$$

This implies that

$$(ad + bc)/bd = (a'd' + b'c')/b'd'.$$

A similar proof shows that $\cdot$ is well defined.

The reader is asked to verify the associative, commutative, and distributive laws for F. The additive identity of F is $0/b$ and the multiplicative identity of F is b/b, where $b \neq 0$. For $a/b \in F$, the additive inverse is

$$(-a)/b = a/(-b)$$

and the multiplicative inverse is b/a (when $a \neq 0$). Thus, F is a field.

We now show that R can be embedded in F. Let

$$R' = \{a/1 \mid a \in R\} \subseteq F.$$

Then R' is a subring of F. Define $f : R \to R'$ by $f(a) = a/1$ for all $a \in R$. Then $a = b$ if and only if $a \cdot 1 = 1 \cdot b$ if and only if $a/1 = b/1$ if and only if $f(a) = f(b)$. Hence, f is a one-one function. Now

$$f(a + b) = (a + b)/1 = (a \cdot 1 + 1 \cdot b)/1 \cdot 1 = a/1 + b/1 = f(a) + f(b)$$

and

$$f(ab) = ab/1 = ab/1 \cdot 1 = a/1 \cdot b/1 = f(a) \cdot f(b).$$

From the definition of f, f is onto R'. Thus, f is an isomorphism of R onto $R' \subseteq F$. ■

The above theorem gives another instance of the power of the concept of an equivalence relation. We have once again used the notion of an ordered pair in a fundamental manner.

Definition 9.1.5 *Let R be an integral domain. A field F is called a **quotient field** of R or a **field of quotients** of R if there exists a subring R_1 of F such that*

(i) $R \simeq R_1$ and

(ii) for all $x \in F$, there exists $a, b \in R_1$ with $b \neq 0$ such that $x = ab^{-1}$.

Let us now show that for the given integral domain R, the field constructed in Theorem 9.1.4 is a quotient field of R. Let $x \in F$. Then $x = a/b$, where $(a, b) \in S$. Now $(a, 1) \in S$ and $(b, 1) \in S$. Thus, $a/1,\ b/1\ \in R'$ and

$$a/b = a/1 \cdot 1/b = (a/1) \cdot (b/1)^{-1}.$$

Hence, F is a quotient field of R. We call F the **quotient field** or the **field of quotients** or R.

Theorem 9.1.6 *Let R be an integral domain and F its field of quotients. Let R' be an integral domain contained in a field K' and set*

$$F' = \{a'(b')^{-1} \mid a', b' \in R', b' \neq 0\}.$$

Then F' is the smallest subfield of K' which contains R' and any isomorphism of R onto R' has a unique extension to an isomorphism of F onto F'.

Proof. By Exercise 2 (page 168), F' is the smallest subfield of K' which contains R'. Let f be an isomorphism of R onto R'. Let $a/b \in F$. If $f(a) = a'$ and $f(b) = b'$, define $g : F \to F'$ by

$$g(a/b) = a'(b')^{-1} = f(a)f(b)^{-1}.$$

Identifying the ring R with the set $\{a/1 \mid a \in R\}$, it is clear that $f = g\,|_R$. Now $a/b = c/d$ if and only if $ad = bc$ if and only if $f(ad) = f(bc)$ if and only if $f(a)f(d) = f(b)f(c)$ if and only if $f(a)f(b)^{-1} = f(c)f(d)^{-1}$ if and only if $g(a/b) = g(c/d)$. Therefore, g is a one-one function. From the definition of g, it follows that g is onto F'. Now

$$\begin{aligned} g(a/b + c/d) &= g((ad + bc)/bd) \\ &= f(ad + bc)(f(bd))^{-1} \\ &= [f(a)f(d) + f(b)f(c)][f(b)^{-1}f(d)^{-1}] \\ &= f(a)f(b)^{-1} + f(c)f(d)^{-1} \\ &= g(a/b) + g(c/d) \end{aligned}$$

and

$$\begin{aligned} g(a/b \cdot c/d) &= g(ac/bd) \\ &= f(ac)(f(bd))^{-1} \\ &= [f(a)f(c)][f(b)^{-1}f(d)^{-1}] \\ &= f(a)f(b)^{-1}f(c)f(d)^{-1} \\ &= g(a/b)g(c/d) \end{aligned}$$

for all $a/b, c/d \in F$. Thus, g is an isomorphism of F onto F'.

Let g' be any other isomorphism of F onto F' such that $f = g' \mid_R$. Then

$$\begin{aligned} g'(a/b) &= g'(a/1 \cdot (b/1)^{-1}) \\ &= g'(a/1)g'((b/1)^{-1}) \\ &= g'(a/1)g'(b/1)^{-1} \\ &= f(a)f(b)^{-1} \\ &= g(a/b) \end{aligned}$$

for all $a/b \in F$, so $g' = g$. Thus, there is a unique extension of f. ■

We can conclude from this result that the field of quotients F of an integral domain R is "the" smallest field containing R in the sense that there does not exist a field K such that $R \subset K \subset F$.

The field F' in Theorem 9.1.6 is called the **quotient field** of R' in K. In view of Theorem 9.1.6 and the comments preceding it, we do not differentiate between the notation a/b and ab^{-1} for the elements of F.

Worked-Out Exercises

♢ **Exercise 1** Let $D = \{\frac{a}{b} \in \mathbb{Q} \mid 5 \text{ does not divide } b\}$. Show that D is a subring of $\mathbb{Q}$ with 1. Find the quotient field of D.

Solution: Let $a/b, c/d \in D$. Because 5 does not divide b and 5 does not divide d, 5 does not divide bd. Thus, $(ad - bc)/bd \in D$ and $ac/bd \in D$. Hence, D is a subring of $\mathbb{Q}$. Also, $1 = 1/1 \in D$. Because $\mathbb{Z} \subseteq D \subseteq \mathbb{Q}$ and $\mathbb{Q}$ is the quotient field of $\mathbb{Z}$, $\mathbb{Q}$ is the quotient field of D.

Exercise 2 Let S be a ring and f a one-one function of S onto a set T. Show that suitable addition and multiplication can be defined on T so that T becomes a ring isomorphic to S under f.

Solution: Define binary operations $+$ and $\cdot$ on T as follows: Let $t_1, t_2 \in T$. Because f maps S onto T, there exist $s_1, s_2 \in S$ such that $f(s_1) = t_1$ and $f(s_2) = t_2$. Define

$$\begin{aligned} t_1 + t_2 &= f(s_1 + s_2) \text{ and} \\ t_1 \cdot t_2 &= f(s_1 s_2). \end{aligned}$$

First we show that both these binary operations are well defined. Let $t_1, t_2, t_3, t_4 \in T$ be such that $t_1 = t_3$ and $t_2 = t_4$. Because f maps S onto T, there exist $s_1, s_2, s_3, s_4 \in S$ such that $f(s_1) = t_1$, $f(s_2) = t_2$, $f(s_3) = t_3$, and $f(s_4) = t_4$. Therefore, $f(s_1) = f(s_3)$ and $f(s_2) = f(s_4)$. Because f is one-one, $s_1 = s_3$ and $s_2 = s_4$. Hence, $t_1 + t_2 = f(s_1 + s_2) = f(s_3 + s_4) = t_3 + t_4$ and $t_1 \cdot t_2 = f(s_1 s_2) = f(s_3 s_4) = t_3 \cdot t_4$. Thus, $+$ and $\cdot$ are well defined. It is now a routine verification to show that $(T, +, \cdot)$ is a ring. We verify some of the properties and leave others as an exercise. First we show that $+$ is associative. Now $t_2 + t_3 = f(s_2 + s_3)$ and $t_1 + t_2 = f(s_1 + s_2)$. Thus, $t_1 + (t_2 + t_3) = f(s_1 + (s_2 + s_3)) = f((s_1 + s_2) + s_3)$ (because $+$ is associative for S) $= (t_1 + t_2) + t_3$. Hence, $+$ is associative for T. Also, $f(0) + t_1 = f(0 + s_1) = f(s_1) = f(s_1 + 0) = t_1 + f(0)$. This implies that $f(0)$ is the additive identity. Similarly, we can verify the other properties of a ring. It is immediate that f is a homomorphism and because f is one-one and f maps S onto T, S is isomorphic to T.

Exercises

1. Prove the associative, commutative, and distributive laws in Theorem 9.1.4.
2. Let R be an integral domain, which is a subring of a field F. Let $F' = \{ab^{-1} \mid a, b \in R, b \neq 0\}$. Show that F' is a subfield of F. Furthermore, show that F' is the smallest subfield of F which contains R.
3. Let R and R' be integral domains contained in fields. Set $F = \{ab^{-1} \mid a, b \in R, b \neq 0\}$ and $F' = \{a'b'^{-1} \mid a', b' \in R', b' \neq 0'\}$. Suppose f is an isomorphism of R onto R'. Prove that f has a unique extension to an isomorphism of F onto F'.
4. Prove that any field R is equal to its field of quotients F in the sense that $f(R) = F$, where f is the isomorphism defined in Theorem 9.1.4.
5. Prove that isomorphic integral domains have isomorphic fields of quotients.
6. Find the field of quotients of the integral domains $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{2}]$.
7. Let R be a ring of characteristic $n > 0$ and

$$R \times \mathbb{Z}_n = \{(r, [m]) \mid r \in R \text{ and } [m] \in \mathbb{Z}_n\}.$$

Define $+$ and $\cdot$ on $R \times \mathbb{Z}_n$ by

$$\begin{aligned}(a,[m])+(b,[t]) &= (a+b,[m+t]),\\ (a,[m])\cdot(b,[t]) &= (ab,[mt])\end{aligned}$$

for all $a, b \in R$, $[m], [t] \in \mathbb{Z}_n$. Prove that

(i) the above two operations are well defined,

(ii) $(R \times \mathbb{Z}_n, +, \cdot)$ is a ring with 1,

(iii) $(R \times \mathbb{Z}_n, +, \cdot)$ is of characteristic n,

(iv) there exists a monomorphism from R into $(R \times \mathbb{Z}_n, +, \cdot)$.

8. Let S and R' be disjoint rings with the property that S contains a subring S' such that there is an isomorphism f' of S' onto R'. Prove that there is a ring R containing R' and an isomorphism f of S onto R such that $f' = f\,|_{S'}$.

Chapter 10

Direct Sum of Rings

In this chapter, we construct some new rings from a given family $\{R_i \mid i \in I\}$ of rings. For this purpose, we introduce the complete direct sum, the direct sum, and the subdirect sum of this family. The results developed in this chapter also help us to obtain structure results of rings.

10.1 Complete Direct Sum and Direct Sum

Let $\{R_i \mid i \in I\}$ be a family of rings indexed by a nonempty set I. The Cartesian product $\Pi\{R_i \mid i \in I\}$ of the sets R_i is the set of all functions $f : I \longrightarrow \cup\{R_i \mid i \in I\}$ such that $f(i) \in R_i$ for all $i \in I$. Let $f, g \in \Pi\{R_i \mid i \in I\}$. Define $f + g$, fg by

$$\begin{aligned}(f+g)(i) &= f(i)+g(i)\\ (fg)(i) &= f(i)g(i)\end{aligned}$$

for all $i \in I$. Then $f + g$, $fg \in \Pi\{R_i \mid i \in I\}$. It can be easily verified that $\Pi\{R_i \mid i \in I\}$ together with the above two operations is a ring. This ring is called the **complete direct sum** of the family of rings $\{R_i \mid i \in I\}$ and is denoted by $\Pi_{i\in I}R_i$. The zero element of $\Pi_{i\in I}R_i$ is the function $0 : I \longrightarrow \cup\{R_i \mid i \in I\}$ defined by $0(i) = 0_i$, the zero element of R_i, for all $i \in I$. The additive inverse of $f \in \Pi_{i\in I}R_i$ is the function $-f : I \longrightarrow \cup\{R_i \mid i \in I\}$ defined by $(-f)(i) = -f(i) \in R_i$ for all $i \in I$. Let $f \in \Pi_{i\in I}R_i$ and let $f(i) = a_i \in R_i$ for all $i \in I$. Usually f is identified with the image set $\{a_i \mid i \in I\}$. Using this notation, the above two operations can be defined by

$$\begin{aligned}\{a_i \mid i \in I\} + \{b_i \mid i \in I\} &= \{a_i + b_i \mid i \in I\}\\ \{a_i \mid i \in I\} \cdot \{b_i \mid i \in I\} &= \{a_i b_i \mid i \in I\}\end{aligned}$$

for all $a_i, b_i \in R_i$ for all $i \in I$.

Suppose now that I is a finite set, say, $I = \{1, 2, \ldots, n\}$. In this case, the complete direct sum is denoted by $\oplus_{i\in I}R_i = R_1 \oplus R_2 \oplus \cdots \oplus R_n$ and an element $\{a_i \mid i \in I\}$ is usually written as an n-tuple $(a_1, a_2, \ldots, a_n)$.

Definition 10.1.1 *The **direct sum** of a family of rings $\{R_i \mid i \in I\}$, denoted by $\oplus_{i\in I}R_i$, is the set*

$$\oplus_{i\in I}R_i = \{\{a_i \mid i \in I\} \in \Pi_{i\in I}R_i \mid a_i \neq 0 \text{ for at most finitely many } i \in I\}.$$

Theorem 10.1.2 *Let $\{R_i \mid i \in I\}$ be a family of rings. Then*
(i) $\oplus_{i\in I}R_i$ is a subring of the complete direct sum of rings $\Pi_{i\in I}R_i$;
(ii) for all $k \in I$, the function $i_k : R_k \to \oplus_{i\in I}R_i$ defined by

$$i_k(a) = \{\{a_i \mid i \in I\} \mid a_i = 0 \text{ for all } i \neq k \text{ and } a_k = a\}$$

for all $a \in R_k$, is a monomorphism of rings;
(iii) for all $k \in I$, $i_k(R_k)$ is an ideal of $\oplus_{i\in I}R_i$.

Proof. (i) Let $\{a_i \mid i \in I\}$ and $\{b_i \mid i \in I\}$ be two elements of $\oplus_{i\in I}R_i$. Because $a_i \neq 0$ for at most finitely many $i \in I$ and $b_i \neq 0$ for at most finitely many $i \in I$, it follows that $a_i - b_i \neq 0$ for at most finitely many $i \in I$ and $a_i b_i \neq 0$ for at most finitely many $i \in I$. Hence, $\{a_i \mid i \in I\} - \{b_i \mid i \in I\} \in \oplus_{i\in I}R_i$ and $\{a_i \mid i \in I\}\{b_i \mid i \in I\} \in \oplus_{i\in I}R_i$. Thus, $\oplus_{i\in I}R_i$ is a subring.

(ii) Let $a, b \in R_k$. Then $i_k(a+b) = \{\{a_i \mid i \in I\} \mid a_i = 0$ for all $i \neq k$ and $a_k = a+b\} = \{\{a'_i \mid i \in I\} \mid a'_k = 0$ for all $i \neq k$ and $a'_i = a\}+ \{\{b'_i \mid i \in I\} \mid b'_i = 0$ for all $i \neq k$ and $b'_k = b\} = i_k(a) + i_k(b)$. Similarly,

$i_k(ab) = i_k(a)i_k(b)$. Thus, i_k is a homomorphism. By the definition of i_k, we find that i_k is one-one. Hence, i_k is a monomorphism.

(iii) Because i_k is a monomorphism, $i_k(R_k)$ is a subring of $\oplus_{i\in I}R_i$. Let $\{b_i \mid i \in I\} \in \oplus_{i\in I}R_i$ and $\{a_i \mid i \in I\} \in i_k(R_k)$. Because $a_i = 0$ for all $i \neq k$, $b_ia_i = 0$ for all $i \neq k$. Also, for $i = k$, $b_k, a_k \in R_k$. Therefore, $b_ka_k \in R_k$. Thus, $\{b_i \mid i \in I\}\{a_i \mid i \in I\} \in i_k(R_k)$, proving that $i_k(R_k)$ is a left ideal. Similarly, $\{a_i \mid i \in I\}\{b_i \mid i \in I\} \in i_k(R_k)$. Hence, $i_k(R_k)$ is an ideal.

■

By Theorem 10.1.2, we find that R_k is isomorphic to the subring $i_k(R_k)$ of $\oplus_{i\in I}R_i$. Identifying R_k with $i_k(R_k)$, we can say that $\oplus_{i\in I}R_i$ contains R_k as an ideal.

Let $I = \{1, 2, \ldots, n\}$ and $\{R_i \mid i \in I\}$ be a finite family of rings. From the definition of direct sum, it follows that the complete direct sum and the direct sum of this family is the same. Hence, by Theorem 10.1.2, we can say that the direct sum, $R_1 \oplus R_2 \oplus \cdots \oplus R_n$, contains each of $R_1, R_2, \ldots, R_n$ as an ideal.

We now investigate the conditions under which a ring R is isomorphic to a direct sum of a family of ideals (considering each ideal as a ring) of R.

Definition 10.1.3 *Let I be a finite nonempty set, say, $\{1, 2, \ldots, n\}$, and $\{A_i \mid i \in I\}$ be a family of ideals of a ring R. Then the sum of this finite family, denoted by $\sum_{i\in I} A_i$, is the set*

$$\sum_{i\in I} A_i = \{a_1 + a_2 + \cdots + a_n \mid a_i \in A_i,\ i = 1, 2,\ \ldots,\ n\}.$$

If I is empty, then let us take $\sum_{i\in I} A_i = \{0\}$.

If $I = \{1, 2, \ldots, n\}$, then we also use the notation $A_1 + A_2 + \cdots + A_n$ to denote the sum $\sum_{i\in I} A_i$.

We leave the proof of the following theorem as an exercise.

Theorem 10.1.4 *Let $\{A_i \mid i \in I\}$ be a finite family of ideals of a ring R. Then*
(i) $\sum_{i\in I} A_i$ is an ideal of R,
(ii) $A_i \subseteq \sum_{j\in I} A_j$ for all $i \in I$,
(iii) if A is an ideal of R such that $A_i \subseteq A$ for all $i \in I$, then $\sum_{i\in I} A_i \subseteq A$. ■

Definition 10.1.5 *Let $\{A_i \mid i \in I\}$ be a family of ideals of a ring R, where I is finite or infinite. Then the sum of this family, denoted by $\sum_{i\in I} A_i$, is the set*

$$\sum_{i\in I} A_i = \{a \in R \mid a \in \sum_{i\in I_0} A_i \text{ for some finite subset } I_0 \text{ of } I\}.$$

Theorem 10.1.6 *Let $\{A_i \mid i \in I\}$ be a family of ideals of a ring R. Then $\sum_{i\in I} A_i$ is an ideal of R.*

Proof. Because $0 \in \sum_{i\in I} A_i$, $\sum_{i\in I} A_i \neq \emptyset$. Let $a, b \in \sum_{i\in I} A_i$ and $r \in R$. Then $a \in \sum_{i\in I_1} A_i$ and $b \in \sum_{i\in I_2} A_i$ for some finite subsets I_1 and I_2 of I. Let $I_3 = I_1 \cup I_2$. Then $a, b \in \sum_{i\in I_3} A_i$. By Theorem 10.1.4, $\sum_{i\in I_3} A_i$ is an ideal of R. Hence, $a - b$, ar, $ra \in \sum_{i\in I_3} A_i$. Thus, $a - b$, ar, $ra \in \sum_{i\in I} A_i$, so $\sum_{i\in I} A_i$ is an ideal of R. ■

Definition 10.1.7 *Let $\{A_i \mid i \in I\}$ be a finite family of ideals of a ring R. A sum $\sum_{i\in I} A_i$ of $\{A_i \mid i \in I\}$ is called a **direct sum** if for all $k \in I$,*

$$A_k \cap \sum_{i\in I,\ i\neq k} A_i = \{0\}.$$

Lemma 10.1.8 *Let $\{A_i \mid i \in I\}$ be a finite family of ideals of a ring R. If $\sum_{i\in I} A_i$ is a direct sum, then for all $a \in A_k$, $b \in A_l$, $k \neq l$, $ab = 0$.*

Proof. Let $a \in A_k$, $b \in A_l$, and $k \neq l$. Because A_k and A_l are ideals, $ab \in A_k$ and $ab \in A_l$. Because $A_l \subseteq \sum_{i\in I,\ i\neq k} A_i$, $ab \in \sum_{i\in I,\ i\neq k} A_i$. Therefore, $ab \in A_k \cap \sum_{i\in I,\ i\neq k} A_i$. Because $\sum_{i\in I} A_i$ is a direct sum, $A_k \cap \sum_{i\in I,\ i\neq k} A_i = \{0\}$. Hence, $ab = 0$. ■

Theorem 10.1.9 *Let $\{A_i \mid i \in I\}$ be a family of ideals of a ring R, $I = \{1, 2, \ldots, n\}$. Then the following conditions are equivalent.*
(i) $\sum_{i\in I} A_i$ is a direct sum.
(ii) $a_1 + a_2 + \cdots + a_n = 0$, $a_i \in A_i$, $i \in I$, implies that $a_i = 0$ for all $i \in I$.

(iii) Each element $a \in \sum_{i\in I} A_i$ is uniquely expressible in the form

$$a = a_1 + a_2 + \cdots + a_n,$$

where $a_i \in A_i$, $i \in I$.

Proof. (i)⇒(ii) Let $a_1 + a_2 + \cdots + a_n = 0$, $a_i \in A_i$, $i \in I$. Let $k \in I$. Now

$$-a_k = a_1 + a_2 + \cdots + a_{k-1} + a_{k+1} + \cdots + a_n \in A_k \cap \sum_{i\in I,\ i\neq k} A_i = \{0\}.$$

Hence, $a_k = 0$.

(ii)⇒(iii) Let $a = a_1 + a_2 + \cdots + a_n = b_1 + b_2 + \cdots + b_n$, where $a_i, b_i \in A_i$ for all $i \in I$. Then $(a_1 - b_1) + (a_2 - b_2) + \cdots + (a_n - b_n) = 0$. Hence, by (ii), $a_i - b_i = 0$ for all $i \in I$, i.e., $a_i = b_i$ for all $i \in I$.

(iii)⇒(i) Let $a \in A_k \cap \sum_{i\in I,\ i\neq k} A_i$. Then there exist $a_i \in A_i$, $i = 1, 2, \ldots, n$, such that

$$a = a_k = a_1 + a_2 + \cdots + a_{k-1} + a_{k+1} + \cdots + a_n.$$

This implies

$$a_1 + a_2 + \cdots + a_{k-1} + (-a_k) + a_{k+1} + \cdots + a_n = 0.$$

Also, $0 + 0 + \cdots + 0 = 0$. Therefore, by (iii), $a_i = 0$ for all $i \in I$ because 0 is uniquely expressible as a sum of elements of A_i. Thus, $A_k \cap \sum_{i\in I,\ i\neq k} A_i = \{0\}$, so $\sum_{i\in I} A_i$ is a direct sum. ■

Definition 10.1.10 *A ring R is said to be an **internal direct sum** of a finite family of ideals $\{A_1, A_2, \ldots, A_n\}$ if*

(i) $R = A_1 + A_2 + \cdots + A_n$ and

(ii) $A_1 + A_2 + \cdots + A_n$ is a direct sum.

Theorem 10.1.11 *Let R be a ring and $\{A_i \mid i \in I\}$ be a finite family of ideals of R. If R is an internal direct sum of $\{A_i \mid i \in I\}$, then*

$$R \simeq \oplus_{i\in I} A_i.$$

Proof. Let $I = \{1, 2, \ldots, n\}$. Suppose R is an internal direct sum of ideals $A_1, A_2, \ldots, A_n$. Let $a \in R$. Then a is uniquely expressible in the form $a = a_1 + a_2 + \cdots + a_n$, where $a_i \in A_i$, $i \in I$. Now $(a_1, a_2, \ldots, a_n) \in \oplus_{i\in I} A_i$. Define $f : R \to \oplus_{i\in I} A_i$ by

$$f(a) = (a_1, a_2, \ldots, a_n).$$

Let $a, b \in R$. Then there exist $a_i, b_i \in A_i$, $i \in I$ such that $a = a_1 + a_2 + \cdots + a_n$ and $b = b_1 + b_2 + \cdots + b_n$. Now $a = b$ if and only if $a_1 + a_2 + \cdots + a_n = b_1 + b_2 + \cdots + b_n$ if and only if $a_i = b_i$ for all $i \in I$ if and only if $(a_1, a_2, \ldots, a_n) = (b_1, b_2, \ldots, b_n)$ if and only if $f(a) = f(b)$. This shows that f is a one-one function. Let $(a_1, a_2, \ldots, a_n) \in \oplus_{i\in I} A_i$. Then $a = a_1 + a_2 + \cdots + a_n \in \sum_{i\in I} A_i = R$ and $f(a) = (a_1, a_2, \ldots, a_n)$. Hence, f is onto $\oplus_{i\in I} A_i$. Finally, we show that f is a homomorphism. Because $a + b = (a_1 + b_1) + (a_2 + b_2) + \cdots + (a_n + b_n)$, we have $f(a + b) = ((a_1 + b_1), (a_2 + b_2), \ldots, (a_n + b_n)) = (a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) = f(a) + f(b)$. By Lemma 10.1.8, for all $i, j \in I$, $i \neq j$, $a_i b_j = 0$. From this, it follows that $ab = a_1b_1 + a_2b_2 + \cdots + a_nb_n$. Thus, $f(ab) = (a_1b_1, a_2b_2, \ldots, a_nb_n) = (a_1, a_2, \ldots, a_n)(b_1, b_2, \ldots, b_n) = f(a)f(b)$. Hence, f is an isomorphism of R onto $\oplus_{i\in I} A_i$, proving that $R \simeq \oplus_{i\in I} A_i$. ■

If R is an internal direct sum of ideals $A_1, A_2, \ldots, A_n$, then we identify R with $\oplus_{i\in I} A_i$ and we usually write

$$R = A_1 \oplus A_2 \oplus \cdots \oplus A_n.$$

Let us now characterize the direct sum of ideals of a ring R with 1 with the help of idempotent elements.

Theorem 10.1.12 *Let R be a ring with 1 and $\{A_1, A_2, \ldots, A_n\}$ be a finite family of ideals of R. Then $R = A_1 \oplus A_2 \oplus \cdots \oplus A_n$ if and only if there exist idempotents $e_i \in A_i$, $i = 1, 2, \ldots, n$, such that*

(i) $1 = e_1 + e_2 + \cdots + e_n$,

(ii) $Re_i = A_i$ for all $i = 1, 2, \ldots, n$, and

(iii) $e_ie_j = e_je_i = 0$ for $i \neq j$.

Proof. Let $R = A_1 \oplus A_2 \oplus \cdots \oplus A_n$. Now $1 \in R$. Thus, there exist $e_i \in A_i$, $i = 1, 2, \ldots, n$, such that $1 = e_1 + e_2 + \cdots + e_n$. Then $e_i = e_1e_i + e_2e_i + \cdots + e_i^2 + \cdots + e_ne_i$. By Lemma 10.1.8, $e_je_i = 0$ for all $j \neq i$. Hence, $e_i = e_i^2$, i.e., e_i is an idempotent for all $i = 1, 2, \ldots, n$. Because $e_i \in A_i$ and A_i is an ideal, $Re_i \subseteq A_i$. Let $a \in A_i$. Then

$$a = a1 = ae_1 + ae_2 + \cdots + ae_n = ae_i \in Re_i$$

because by Lemma 10.1.8, $ae_j = 0$ for all $j \neq i$. Thus, $A_i \subseteq Re_i$. Therefore, we find that $Re_i = A_i$.

Conversely, assume that there exist idempotents $e_i \in A_i$, $i = 1, 2, \ldots, n$, satisfying the given conditions. Let $a \in R$. Then $a = a1 = a(e_1+e_2+\cdots+e_n) = ae_1+ae_2+\cdots+ae_n \in Re_1+Re_2+\cdots+Re_n \subseteq A_1+A_2+\cdots+A_n$. Hence, $R = A_1+A_2+\cdots+A_n$. Let us now show that this sum is direct. Let $a \in A_i \cap (A_1+A_2+\cdots+A_{i-1}+A_{i+1}+\cdots+A_n)$. Then there exist $a_1, a_2, \ldots, a_n \in R$ such that $a_ie_i = a = a_1e_1 + \cdots + a_{i-1}e_{i-1} + a_{i+1}e_{i+1} + \cdots + a_ne_n$. Thus, $a = a_ie_i$ implies that $ae_i = a_ie_i^2 = a_ie_i = a$ and $a = a_1e_1 + \cdots + a_{i-1}e_{i-1} + a_{i+1}e_{i+1} + \cdots + a_ne_n$ implies that $ae_i = a_1e_1e_i + \cdots + a_{i-1}e_{i-1}e_i + a_{i+1}e_{i+1}e_i + \cdots + a_ne_ne_i = a0 + \cdots + a0 = 0$ (because by (iii), $e_ie_j = 0$ for $i \neq j$). Hence, $a = 0$, proving that $R = A_1 \oplus A_2 \oplus \cdots \oplus A_n$. ■

Let us now consider another type of subring of the complete direct sum $\Pi_{i\in I}R_i$ of a family of rings $\{R_i \mid i \in I\}$. For this, let us note that the mapping $\pi_k : \Pi_{i\in I}R_i \longrightarrow R_k$ defined by

$$\pi_k(\{a_i \mid i \in I\}) = a_k$$

is an epimorphism of the ring $\Pi_{i\in I}R_i$ onto the ring R_k. π_k is called the kth **canonical projection.**

Definition 10.1.13 *A subring T of $\Pi_{i\in I}R_i$ is called a* ***subdirect sum*** *of the family of rings $\{R_i \mid i \in I\}$ if $\pi_i \mid_T$ (the restriction of π_i to T) is an epimorphism of T onto R_i. We denote T by $\oplus^s_{i\in I}R_i$.*

Theorem 10.1.14 *A ring S is isomorphic to a subdirect sum of a family $\{R_i \mid i \in I\}$ of rings if and only if S contains a family of ideals $\{A_i \mid i \in I\}$ such that $\cap_{i\in I}A_i = \{0\}$.*

Proof. Suppose S is isomorphic to a subdirect sum of a family $\{R_i \mid i \in I\}$ of rings. Then there exists a subring T of $\Pi_{i\in I}R_i$ such that $S \simeq T$ and $T = \oplus^s_{i\in I}R_i$. Let α be the isomorphism of S onto T. Then $\pi_i\alpha : S \longrightarrow R_i$ is an epimorphism. Let $A_i = \text{Ker } \pi_i\alpha$. Then A_i is an ideal of S. Let $a \in \cap_{i\in I}A_i$. Then $(\pi_i\alpha)(a) = 0$ for all $i \in I$. Thus, $\pi_i(\alpha(a)) = 0$, i.e., the ith component of $\alpha(a)$ is 0 for all $i \in I$. Hence, $\alpha(a) = 0$. Because α is one-one, $a = 0$. This proves that $\cap_{i\in I}A_i = \{0\}$.

Conversely, suppose S contains a family of ideals $\{A_i \mid i \in I\}$ such that $\cap_{i\in I}A_i = \{0\}$. Consider the family $\{S/A_i \mid i \in I\}$ of quotient rings. Let $R = \Pi_{i\in I}S/A_i$. Define $\beta : S \longrightarrow R$ by

$$\beta(a) = \{a + A_i \mid i \in I\}$$

for all $a \in S$. Then β is a homomorphism. Let $a \in S$. Now $a \in \text{Ker } \beta$ if and only if $\beta(a) = 0$ if and only if $a + A_i = 0$ for all $i \in I$ if and only if $a \in A_i$ for all $i \in I$ if and only if $a \in \cap_{i\in I}A_i$ if and only if $a = 0$. Therefore, Ker $\beta = \{0\}$. Thus, β is a monomorphism. Let $\beta(S) = T$. Then T is a subring of R and also $\pi_i \mid_T$ is an epimorphism. ■

Worked-Out Exercises

♢ **Exercise 1** An idempotent e of a ring R is called a **central idempotent** if $e \in C(R)$.

Let R be a ring with 1 and e be a central idempotent in R. Show that

(a) $1 - e$ is a central idempotent in R;

(b) eR and $(1 - e)R$ are ideals of R;

(c) $R = eR \oplus (1 - e)R$.

Solution: (a) $(1-e)(1-e) = 1-e-e+e^2 = 1-e-e+e = 1-e$. Also, for all $a \in R$, $a(1-e) = a-ae = a-ea = (1-e)a$. Hence, $1 - e$ is a central idempotent.

(b) Now eR is a right ideal of R. Let $a \in R$. Then $a(eR) = (ae)R = (ea)R$ (because $e \in C(R)) = e(aR) \subseteq eR$. Hence, eR is also a left ideal. Thus, eR is an ideal of R. Similarly, $(1 - e)R$ is an ideal of R.

(c) Let $a \in R$. Then $a = ea+a-ea = ea+(1-e)a \in eR+(1-e)R$. Hence, $R = eR+(1-e)R$. Suppose $b \in eR \cap (1-e)R$. Then there exist $c, d \in R$ such that $b = ec = (1-e)d$. Hence, $eb = e^2c = ec = b$ and $eb = e(1-e)d = (e-e^2)d = (e-e)d = 0$. Thus, $b = 0$. As a result, $R = eR \oplus (1-e)R$.

♢ **Exercise 2** Let A and B be two ideals of a ring R such that $R = A \oplus B$. Show that $R/A \simeq B$ and $R/B \simeq A$.

Solution: Let $x \in R$. Then x can be uniquely expressed as $x = a + b$, where $a \in A$ and $b \in B$. Define $f : R \to B$ by $f(x) = b$. Clearly f is well defined. Let $b \in B$. Then $b = 0 + b \in A + B$. Hence, $f(b) = b$, which shows that f is onto B. Let $x, y \in R$. Then there exist $a_1, a_2 \in A$ and $b_1, b_2 \in B$ such that $x = a_1 + b_1$ and $y = a_2 + b_2$. Now $x + y = a_1 + b_1 + a_2 + b_2 = (a_1 + a_2) + (b_1 + b_2) \in A + B$ and $xy = (a_1 + b_1)(a_2 + b_2) = a_1a_2 + a_1b_2 + b_1a_2 + b_1b_2$. Because $a_1b_2, b_1a_2 \in A \cap B$ and $A \cap B = \{0\}$, $a_1b_2 = 0$ and $b_1a_2 = 0$. Therefore, $xy = a_1a_2 + b_1b_2 \in A + B$. Hence, $f(x + y) = b_1 + b_2 = f(x) + f(y)$ and $f(xy) = b_1b_2 = f(x)f(y)$. Thus, f is an epimorphism. Therefore, by the first isomorphism theorem (Theorem 8.3.14), $R/\text{Ker} f \simeq B$. Let $x \in \text{Ker } f$. Then $f(x) = 0$. Because $x \in \text{Ker } f \subseteq R$, there exist $a \in A$ and $b \in B$ such that $x = a + b$. Now $f(x) = b$ and this implies that $b = 0$. Therefore, $x = a \in A$, so $\text{Ker } f \subseteq A$. On the other hand, let $a \in A$. Then $a = a + 0 \in A + B$. Therefore, $f(a) = 0$, so $a \in \text{Ker } f$. Thus, $A \subseteq \text{Ker } f$. Hence, $A = \text{Ker } f$, so $R/A \simeq B$. Similarly, $R/B \simeq A$.

Exercise 3 Let $R = R_1 \oplus R_2 \oplus \cdots \oplus R_n$ be the direct of sum of rings $R_1, R_2, \ldots, R_n$ and $1 \in R$. Show that an element $a = (a_1, a_2, \ldots, a_n) \in R$ is a unit if and only if a_i is a unit in R_i for all $i = 1, 2, \ldots, n$.

Solution: Because $1 \in R = R_1 \oplus R_2 \oplus \cdots \oplus R_n$, $1 = (e_1, e_2, \ldots, e_n)$, where e_i is the identity of R_i for all $i = 1, 2, \ldots, n$. Suppose $a = (a_1, a_2, \ldots, a_n) \in R$ is a unit. Then there exists $b = (b_1, b_2, \ldots, b_n) \in R$ such that $ab = 1 = ba$. Thus, $(a_1, a_2, \ldots, a_n)(b_1, b_2, \ldots, b_n) = (e_1, e_2, \ldots, e_n) = (b_1, b_2, \ldots, b_n)(a_1, a_2, \ldots, a_n)$. From this, it follows that $a_ib_i = e_i = b_ia_i$ for all $i = 1, 2, \ldots, n$. Hence, a_i is a unit in R_i for all $i = 1, 2, \ldots, n$. Conversely, assume that a_i is a unit in R_i for all $i = 1, 2, \ldots, n$. Thus, there exists $b_i \in R_i$ such that $a_ib_i = e_i = b_ia_i$ for all $i = 1, 2, \ldots, n$. Let $b = (b_1, b_2, \ldots, b_n)$. Then $ab = 1 = ba$, proving that a is a unit.

♢ **Exercise 4** Let R be a direct of sum of rings $R_1, R_2, \ldots, R_n$ with identity. Let A be an ideal of R. Show that there exist ideals A_i in R_i, $i = 1, 2, \ldots, n$, such that $A = A_1 \oplus A_2 \oplus \cdots \oplus A_n$.

Solution: For all k, $1 \leq k \leq n$, define $\alpha_k : \oplus R_i \to R_k$ by

$$\alpha_k((a_1, a_2, \ldots, a_n)) = a_k$$

for all $(a_1, a_2, \ldots, a_n) \in \oplus R_i$. It can be easily verified that α_k is an epimorphism. Let $\alpha_k(A) = A_k$. Then A_k is an ideal of R_k. We now show that $A = A_1 \oplus A_2 \oplus \cdots \oplus A_n$. Let $a = (a_1, a_2, \ldots, a_n) \in A$. Now $\alpha_k(a) = a_k \in A_k$. Therefore, $a \in A_1 \oplus A_2 \oplus \cdots \oplus A_n$, so $A \subseteq A_1 \oplus A_2 \oplus \cdots \oplus A_n$. Suppose now that $b = (b_1, b_2, \ldots, b_n) \in A_1 \oplus A_2 \oplus \cdots \oplus A_n$. Then $b_k \in A_k = \alpha_k(A)$. Therefore, there exists an element $a = (a_1, a_2, \ldots, a_{k-1}, b_k, a_{k+1}, \ldots, a_n) \in A$. Now $(0, 0, \ldots, 0, b_k, 0, \ldots, 0) = (0, 0, \ldots, 1, \ldots, 0)(a_1, a_2, \ldots, a_{k-1}, b_k, a_{k+1}, \ldots, a_n) \in A$ for all $k = 1, 2, \ldots, n$. Hence, $(b_1, b_2, \ldots, b_n) = (b_1, 0, \ldots, 0) + (0, b_2, \ldots, 0) + \cdots + (0, 0, \ldots, b_n) \in A$ showing that $A_1 \oplus A_2 \oplus \cdots \oplus A_n \subseteq A$. Thus, $A = A_1 \oplus A_2 \oplus \cdots \oplus A_n$.

♢ **Exercise 5** Let R be a ring with 1. Suppose that A and B are ideals of R such that $R = A + B$. Show that

$$R/(A \cap B) \simeq R/A \oplus R/B.$$

(This result is known as the **Chinese remainder theorem for rings**.)

Solution: Define $f : R \to R/A \oplus R/B$ by

$$f(x) = (x + A, x + B)$$

for all $x \in R$. Let $x, y \in R$. Then

$$\begin{aligned} f(x + y) &= ((x + y) + A, (x + y) + B) \\ &= ((x + A) + (y + A), (x + B) + (y + B)) \\ &= (x + A, x + B) + (y + A, y + B) \\ &= f(x) + f(y). \end{aligned}$$

Similarly, $f(xy) = f(x)f(y)$. Hence, f is a homomorphism. Now $R = A + B$ implies that $1 = a + b$ for some $a \in A$ and $b \in B$. Thus, $a + B = (1 - b) + B = (1 + B) + (-b + B) = 1 + B$ because $-b \in B$. Similarly, $b + A = 1 + A$. Let $(x + A, y + B) \in R/A \oplus R/B$. Now $xb + ya \in R$. Therefore,

$$\begin{aligned} f(xb + ya) &= ((xb + ya) + A, (xb + ya) + B) \\ &= ((xb + A) + (ya + A), (xb + B) + (ya + B)) \\ &= ((xb + A) + (0 + A), (0 + B) + (ya + B)) \text{ (because } a \in A,\ b \in B) \\ &= ((xb + A)\ , (ya + B)) \\ &= ((x + A)(b + A), (y + B)(a + B)) \\ &= ((x + A)(1 + A), (y + B)(1 + B)) \\ &= (\ x + A\ ,\ y + B\ ). \end{aligned}$$

Hence, f is an epimorphism. By the first isomorphism theorem (Theorem 8.3.14),

$$R/\text{Ker } f \simeq R/A \oplus R/B.$$

We now show that Ker $f = A \cap B$.

$$\begin{aligned}
\text{Ker } f &= \{x \in R \mid f(x) = 0\} \\
&= \{x \in R \mid (x + A, x + B) = (A, B)\} \\
&= \{x \in R \mid x + A = A \text{ and } x + B = B\} \\
&= \{x \in R \mid x \in A \text{ and } x \in B\} \\
&= \{x \in R \mid x \in A \cap B\} \\
&= A \cap B.
\end{aligned}$$

Consequently, $R/(A \cap B) \simeq R/A \oplus R/B$.

Exercises

1. Let $R = R_1 \oplus R_2 \oplus \cdots \oplus R_n$ be a direct sum of rings. If A_i is an ideal of R_i, $(1 \leq i \leq n)$, prove that $A = A_1 \oplus A_2 \oplus \cdots \oplus A_n$ is an ideal of R.
2. Let R be a direct of sum of rings $R_1, R_2, \ldots, R_n$ with 1. Let A be an ideal of R. Show that there exist ideals A_i of R_i, $i = 1, 2, \ldots, n$, such that $A = A_1 \oplus A_2 \oplus \cdots \oplus A_n$ and
$$R/A \simeq R_1/A_1 \oplus R_2/A_2 \oplus \cdots \oplus R_n/A_n.$$
3. Show that the ring $\mathbb{Z}$ cannot be expressed as a direct sum of a finite family of proper ideals of $\mathbb{Z}$.
4. If m and n are two positive integers such that $\gcd(m, n) = 1$, prove that $\mathbb{Z}_{mn} \simeq \mathbb{Z}_m \oplus \mathbb{Z}_n$.

Chapter 11

Polynomial Rings

The study of polynomials dates back to 1650 B.C., when Egyptians were solving certain linear polynomial equations. In 600 B.C., Hindus had learned how to solve quadratic equations. However, polynomials, as we know them today, i.e., polynomials written in our notation, did not exist until approximately 1700 A.D.

About 400 A.D., the use of symbolic algebra began to appear in India and Arabia. Some mark the use of symbols in algebra as the first level of abstraction in mathematics.

11.1 Polynomial Rings

An important class of rings is the so-called class of polynomial rings. We are all familiar with polynomials. We may be used to thinking of a polynomial as an expression of the form $a_0 + a_1x + \cdots + a_nx^n$, where x is a symbol and the a_i are possibly real numbers, or as a function $f(x) = a_0 + a_1x + \cdots + a_nx^n$. However, does one really know what a polynomial is? What really is the symbol x? Why are two polynomials $a_0 + a_1x + \cdots + a_nx^n$ and $b_0 + b_1x + \cdots + b_mx^m$ equal if and only if $n = m$ and $a_i = b_i$, $i = 1, 2, \ldots, n$? In this section, we answer these questions and give some basic properties of polynomials.

Definition 11.1.1 *For any ring R, let $R[x]$ denote the set of all infinite sequences $(a_0, a_1, a_2, \ldots)$, where $a_i \in R$, $i = 0, 1, 2, \ldots$, and where there is a nonnegative integer n (dependent on $(a_0, a_1, a_2, \ldots)$) such that for all integers $k \geq n$, $a_k = 0$. The elements of $R[x]$ are called **polynomials** over R.*

We now define addition and multiplication on $R[x]$ as follows:

$$\begin{aligned}(a_0, a_1, a_2, \ldots) + (b_0, b_1, b_2, \ldots) &= (a_0 + b_0, a_1 + b_1, a_2 + b_2, \ldots)\\ (a_0, a_1, a_2, \ldots) \cdot (b_0, b_1, b_2, \ldots) &= (c_0, c_1, c_2, \ldots),\end{aligned}$$

where

$$c_j = \sum_{i=0}^{j} a_i b_{j-i} \text{ for } \ j = 0, 1, 2, \ldots$$

We leave it to the reader to verify that $(R[x], +, \cdot)$ is a ring. We do note that $(0, 0, \ldots)$ is the additive identity of $R[x]$ and that the additive inverse of $(a_0, a_1, \ldots)$ is $(-a_0, -a_1, \ldots)$. The ring $R[x]$ is called a **ring of polynomials** or a **polynomial ring** over R. It is clear that $R[x]$ is commutative when R is commutative. Also, if R has an identity 1, then $R[x]$ has an identity, namely, $(1, 0, 0, 0, \ldots)$.

The mapping $a \to (a, 0, 0, \ldots)$ is a monomorphism of R into $R[x]$. Thus, R is embedded in $R[x]$. Therefore, we can consider R as a subring of $R[x]$ and we no longer distinguish between a and $(a, 0, 0, \ldots)$.

We now convert our notation of polynomials into a notation which is more familiar to the reader.

Let

$$\begin{aligned}a = ax^0 &\text{ denote } (a, 0, 0, \ldots)\\ ax = ax^1 &\text{denote } (0, a, 0, \ldots)\\ ax^2 &\text{ denote } (0, 0, a, \ldots)\\ &\vdots\end{aligned}$$

Then

$(a_0, a_1, a_2, \ldots, a_n, 0, \ldots) = (a_0, 0, 0, \ldots) + (0, a_1, 0, 0, \ldots) + \cdots + (0, \ldots, 0, a_n, 0, \ldots) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$.

The symbol x is called an **indeterminate** over R and the elements a_0, a_1, $\ldots$, a_n of R are called the **coefficients** of $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$.

The reason two polynomials $a_0 + a_1x + \cdots + a_nx^n$ and $b_0 + b_1x + \cdots + b_mx^m$ are equal if and only if $n = m$ and $a_i = b_i$, $i = 1, 2, \ldots, n$, is that the two sequences $(a_0, a_1, \ldots)$ and $(b_0, b_1, \ldots)$ are equal if and only if $a_i = b_i$, $i = 1, 2, \ldots$. (One must recall that an infinite sequence of elements of R is a function from the set of nonnegative integers into R. Consequently, the concept of an ordered pair is again being used to give a rigorous definition of a mathematical concept.)

If R has an identity 1, then we can consider x an element of $R[x]$. We do this by identifying $1x$ with x, i.e., $(0, 1, 0, \ldots)$ is called x.

The reader can check that the definitions of addition and multiplication of two polynomials are the familiar ones. Thus, when R has an identity, $ax = (a, 0, 0, \ldots)(0, 1, 0, \ldots) = (0, a, 0, \ldots) = (0, 1, 0, \ldots)(a, 0, 0, \ldots) = xa$.

Theorem 11.1.2 *(i) If R is a commutative ring with 1, then $R[x]$ is a commutative ring with 1.*
(ii) If R is an integral domain, then $R[x]$ is also an integral domain.

Proof. (i) Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ be two elements in $R[x]$. Let $f(x)g(x) = c_0 + c_1x + \cdots + c_tx^t$ and $g(x)f(x) = d_0 + d_1x + \cdots + d_sx^s$. Now $c_j = \sum_{i=0}^{j} a_ib_{j-i}$ and $d_j = \sum_{i=0}^{j} b_ia_{j-i}$. Because R is commutative, $c_j = a_0b_j + a_1b_{j-1} + \cdots + a_jb_0 = b_0a_j + b_1a_{j-1} + \cdots + b_ja_0 = d_j$ for all $j = 0, 1, 2, \ldots$. Thus, $R[x]$ is a commutative ring. Because $1 \in R$, $1 \in R[x]$ and $1f(x) = f(x)1 = f(x)$ for all $f(x) \in R[x]$. Hence, $R[x]$ is a commutative ring with 1.

(ii) Let R be an integral domain. Then by (i), $R[x]$ is a commutative ring with 1. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ be two nonzero polynomials in $R[x]$. Then there exist a_i and b_j such that $a_i \neq 0$, $b_j \neq 0$, $a_{i+t} = 0$, and $b_{j+t} = 0$ for all $t \geq 1$. Consider the polynomial

$$f(x)g(x) = c_0 + c_1x + \cdots + c_{n+m}x^{n+m}.$$

Now

$$c_{i+j} = a_0b_{i+j} + a_1b_{i+j-1} + \cdots + a_ib_j + \cdots + a_{i+j}b_0 = a_ib_j \neq 0$$

because R is an integral domain. This implies that $f(x)g(x) \neq 0$. Thus, $R[x]$ is an integral domain. ■

Definition 11.1.3 *Let R be a ring. If $f(x) = a_0 + a_1x + \cdots + a_nx^n$, $a_n \neq 0$, is a polynomial in $R[x]$, then n is called the **degree** of $f(x)$, written* $\deg f(x)$, *and a_n is called the **leading coefficient** of $f(x)$. If R has an identity and $a_n = 1$, then $f(x)$ is called a **monic polynomial**.*

The polynomials of degree 0 in $R[x]$ are exactly those elements from $R\backslash\{0\}$. $0 \in R[x]$ has no degree. We call the elements of R **scalar** or **constant polynomials**.

Theorem 11.1.4 *Let $R[x]$ be a polynomial ring and $f(x)$, $g(x)$ be two nonzero polynomials in $R[x]$.*
(i) If $f(x)g(x) \neq 0$, then $\deg f(x)g(x) \leq \deg f(x) + \deg g(x)$.
(ii) If $f(x) + g(x) \neq 0$, then

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

Proof. (i) If $f(x) = a_0 + a_1x + \cdots + a_nx^n$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$, then

$$f(x)g(x) = a_0b_0 + (a_0b_1 + a_1b_0)x + \cdots + a_nb_mx^{n+m}.$$

If $f(x)g(x) \neq 0$, then at least one of the coefficients of $f(x)g(x)$ is nonzero. If $a_nb_m \neq 0$, then

$$\deg(f(x)g(x)) = n + m = \deg f(x) + \deg g(x).$$

If $a_nb_m = 0$ (which can hold if R has zero divisors), then $\deg(f(x)g(x)) < \deg f(x) + \deg g(x)$.

(ii) If $\deg f(x) > \deg g(x)$, then $\deg(f(x) + g(x)) = \max\{\deg f(x), \deg g(x)\}$. If $\deg f(x) = \deg g(x)$, then it is possible that $f(x) + g(x) = 0$ or $\deg(f(x) + g(x)) < \max\{\deg f(x), \deg g(x)\}$. We leave the details as an exercise. ■

From the proof of Theorem 11.1.4(i), it is immediate that if R is an integral domain, then equality holds in (i).

Example 11.1.5 *Consider the polynomial ring* $\mathbb{Z}_6[x]$. *Let* $f(x) = [1] + [2]x^2$ *and* $g(x) = [1] + [3]x$. *Then*

$$f(x)g(x) = [1] + [3]x + [2]x^2.$$

Hence, $\deg\,(f(x)g(x)) = 2 < 3 = \deg f(x) + \deg g(x)$. *Let* $h(x) = [5] + [4]x^2$. *Then*

$$f(x) + h(x) = [6] + [6]x^2 = [0],$$

so $\deg(f(x) + h(x))$ *is not defined.*

Theorem 11.1.6 (Division Algorithm) *Let* R *be a commutative ring with* 1 *and* $f(x)$, $g(x)$ *be polynomials in* $R[x]$ *with the leading coefficient of* $g(x)$ *a unit in* R. *Then there exist unique polynomials* $q(x)$, $r(x) \in R[x]$ *such that*

$$f(x) = q(x)g(x) + r(x),$$

where either $r(x) = 0$ *or* $\deg r(x) < \deg\, g(x)$.

Proof. If $f(x) = 0$ or $\deg f(x) < \deg g(x)$, then we take $q(x) = 0$ and $r(x) = f(x)$. We now assume that $\deg f(x) \geq \deg g(x)$ and prove the result by induction on $\deg f(x) = n$. If $\deg f(x) = \deg g(x) = 0$, then we have $q(x) = f(x)g(x)^{-1}$ and $r(x) = 0$. Make the induction hypothesis that the theorem is true for all polynomials of degree less than n. Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ have degree n and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ have degree m, where $n \geq m$. The polynomial

$$f_1(x) = f(x) - (a_nb_m^{-1})x^{n-m}g(x) \tag{11.1}$$

has degree less than n because the coefficient of x^n is $a_n - (a_nb_m^{-1})b_m = 0$. Hence, by the induction hypothesis, there exist polynomials $q_1(x)$, $r_1(x) \in R[x]$ such that

$$f_1(x) = q_1(x)g(x) + r_1(x), \tag{11.2}$$

where $r_1(x) = 0$ or $\deg r_1(x) < \deg g(x)$. Substituting the representation of $f_1(x)$ in Eq. (11.2) into Eq. (11.1) and solving for $f(x)$, we obtain

$$f(x) = (q_1(x) + a_nb_m^{-1}x^{n-m})g(x) + r_1(x) = q(x)g(x) + r(x),$$

where $q(x) = q_1(x) + a_nb_m^{-1}x^{n-m}$ and $r(x) = r_1(x)$, the desired representation when $f(x)$ has degree n.

The uniqueness of $q(x)$ and $r(x)$ remains to be shown. Suppose there are polynomials $q'(x)$ and $r'(x) \in R[x]$ such that

$$f(x) = q(x)g(x) + r(x) = q'(x)g(x) + r'(x),$$

where $r(x) = 0$ or $\deg r(x) < \deg g(x)$, $r'(x) = 0$ or $\deg r'(x) < \deg g(x)$. Then

$$r(x) - r'(x) = (q'(x) - q(x))g(x).$$

Suppose $r(x) - r'(x) \neq 0$. Because the leading coefficient of $g(x)$ is a unit,

$$\deg((q'(x) - q(x))g(x)) = \deg(q'(x) - q(x)) + \deg g(x) \geq \deg g(x).$$

This implies that

$$\deg(r(x) - r'(x)) \geq \deg g(x),$$

which is impossible because $\deg r(x)$, $\deg r'(x) < \deg g(x)$. Thus,

$$r(x) - r'(x) = 0 \text{ or } r(x) = r'(x).$$

Therefore,

$$0 = (q'(x) - q(x))g(x). \tag{11.3}$$

Because b_m is a unit, $\deg(((q'(x) - q(x))g(x)) \geq 0$ unless $q'(x) - q(x) = 0$. Thus, from Eq. (11.3), we see that $q'(x) - q(x) = 0$ must be the case. ■

The polynomials $q(x)$ and $r(x)$ in Theorem 11.1.6 are called the **quotient** and **remainder**, respectively, on division of $f(x)$ by $g(x)$.

Definition 11.1.7 *Let* R *be a commutative ring with* 1 *and* $f(x) = a_0 + a_1x + \cdots + a_nx^n \in R[x]$. *For all* $r \in R$, *define*

$$f(r) = a_0 + a_1r + \cdots + a_nr^n.$$

When $f(r) = 0$, *we call* r *a* ***root*** *or* ***zero*** *of* $f(x)$.

In Definition 11.1.7, we think of substituting r for x in $f(x)$. The student is used to doing this freely. However, certain difficulties arise when R is not commutative. For instance, let $f(x) = a - x$, $g(x) = b - x$. Set $h(x) = f(x)g(x)$. Then

$$h(x) = (a - x)(b - x) = ab - (a + b)x + x^2.$$

For $c \in R$,

$$h(c) = ab - (a + b)c + c^2 = ab - ac - bc + c^2$$

while

$$f(c)g(c) = (a - c)(b - c) = ab - cb - ac + c^2.$$

Hence, we cannot draw the conclusion that $h(c) = f(c)g(c)$. However, if R is commutative (with identity), then we can conclude that $h(c) = f(c)g(c)$. Clearly if $k(x) = f(x) + g(x)$, then $k(c) = f(c) + g(c)$.

Definition 11.1.8 *Let R be a commutative ring with 1 and $f(x)$, $g(x) \in R[x]$ be such that $g(x) \neq 0$. We say that $g(x)$ **divides** $f(x)$ or that $g(x)$ is a **factor** of $f(x)$, and write $g(x) \mid f(x)$ if there exists $q(x) \in R[x]$ such that $f(x) = q(x)g(x)$.*

Theorem 11.1.9 (Remainder Theorem) *Let R be a commutative ring with identity. For $f(x) \in R[x]$ and $a \in R$, there exists $q(x) \in R[x]$ such that*

$$f(x) = (x - a)q(x) + f(a).$$

Proof. By applying the division algorithm with $x - a = g(x)$, there exist unique $q(x)$, $r(x) \in R[x]$ such that $f(x) = (x - a)q(x) + r(x)$, where $r(x) = 0$ or $\deg r(x) < 1$. Hence, $r(x)$ is a constant polynomial, say, $r(x) = d$. By substituting a for x, we obtain $f(a) = (a - a)q(a) + d = d$, which yields the desired result. ■

Corollary 11.1.10 (Factorization Theorem) *Let R be a commutative ring with identity. For $f(x) \in R[x]$ and $a \in R$, $x - a$ divides $f(x)$ if and only if a is a root of $f(x)$.*

Proof. Suppose $(x - a) \mid f(x)$. Then there exists $q(x) \in R[x]$ such that $f(x) = (x - a)q(x)$. Hence, $f(a) = (a - a)q(a) = 0$, so a is a root of $f(x)$. Conversely, suppose a is a root of $f(x)$. Then by the remainder theorem (Theorem 11.1.9) and the fact that $f(a) = 0$, we have $f(x) = (x - a)q(x)$. Consequently, $(x - a) \mid f(x)$. ■

Theorem 11.1.11 *Let R be an integral domain and $f(x)$ be a nonzero polynomial in $R[x]$ of degree n. Then $f(x)$ has at most n roots in R.*

Proof. If $\deg f(x) = 0$, then $f(x)$ is a constant polynomial, say, $f(x) = c \neq 0$. Clearly c has no roots in R. Assume that the theorem is true for all polynomials of degree less than n, where $n > 0$ (the induction hypothesis). Suppose $\deg f(x) = n$. If $f(x)$ has no roots in R, then the theorem is true. Suppose $r \in R$ is a root of $f(x)$. Then by Corollary 11.1.10, $f(x) = (x - r)q(x)$, where $\deg q(x) = n - 1$. If there exists any other root $r' \in R$ of $f(x)$, then $0 = f(r') = (r' - r)q(r')$. Because $r' \neq r$ and R is an integral domain, $q(r') = 0$, so r' is a root of $q(x)$. Therefore, any other root of $f(x)$ is also a root of $q(x)$. Because $f(x) = (x - r)q(x)$, any root of $q(x)$ is also a root of $f(x)$. By the induction hypothesis and the fact that $\deg q(x) = n - 1$, there are at most $n - 1$ of these other roots r'. Hence, in all, $f(x)$ has at most n roots in R. ■

We now extend the definition of a polynomial ring from one indeterminate to several indeterminates.

Definition 11.1.12 *For any ring R, we define recursively*

$$R[x_1, x_2, \ldots, x_n] = R[x_1, x_2, \ldots, x_{n-1}][x_n],$$

*where x_1 is an indeterminate over R and x_n is an indeterminate over $R[x_1, x_2, \ldots, x_{n-1}]$. $R[x_1, x_2, \ldots, x_n]$ is called a **polynomial ring in n indeterminates**.*

Before describing the ring $R[x_1, x_2, \ldots, x_n]$, we introduce some notation. We write $\sum_{i_n, \ldots, i_1} r_{i_1 \ldots i_n} x_1^{i_1} \cdots x_n^{i_n}$ for $\sum_{i_n=0}^{k_n} \cdots \sum_{i_1=0}^{k_1} r_{i_1 \ldots i_n} x_1^{i_1} \cdots x_n^{i_n}$, where each $r_{i_1 \ldots i_n} \in R$ and $k_1, \ldots, k_n$ are nonnegative integers.

The ring

$$R[x_1, x_2, \ldots, x_n] = \{ \sum_{i_n, \ldots, i_1} r_{i_1 \ldots i_n} x_1^{i_1} \cdots x_n^{i_n} \mid r_{i_1 \ldots i_n} \in R \}.$$

We have for $n = 2$ that

$$R[x_1, x_2] = R[x_1][x_2] = \{ \sum_{i_2} s_{i_2} x_2^{i_2} \mid s_{i_2} \in R[x_1] \}.$$

Now each s_{i_2} has the form $\sum_{i_1} r_{i_1 i_2} x_1^{i_1}$.
Thus,

$$\begin{aligned} R[x_1, x_2] &= \{\textstyle\sum_{i_2}(\sum_{i_1} r_{i_1 i_2} x_1^{i_1}) x_2^{i_2} \mid r_{i_1 i_2} \in R\} \\ &= \{\textstyle\sum_{i_2} \sum_{i_1} r_{i_1 i_2} x_1^{i_1} x_2^{i_2} \mid r_{i_1 i_2} \in R\} \\ &= \{\textstyle\sum_{i_2, i_1} r_{i_1 i_2} x_1^{i_1} x_2^{i_2} \mid r_{i_1 i_2} \in R.\}. \end{aligned}$$

Definition 11.1.13 *Let R be a subring of the ring S. Let $c_1, c_2, \ldots, c_n$ be elements of S. Define $R[c_1] = \{\sum_i r_i c_1^i \mid r_i \in R\}$ and*

$$R[c_1, c_2, \ldots, c_n] = R[c_1, c_2, \ldots, c_{n-1}][c_n].$$

*We say that $c_1, c_2, \ldots, c_n$ are **algebraically independent** over R if*

$$\sum_{i_n, .., i_1} r_{i_1 \ldots i_n} c_1^{i_1} \ldots c_n^{i_n} = 0$$

can occur only when each $r_{i_1 \ldots i_n} = 0$, where $r_{i_1 \ldots i_n} \in R$.

$R[c_1, c_2, \ldots, c_n]$ is a subring of S and equals the set of all finite sums of the form

$$\sum_{i_n, .., i_1} r_{i_1 \ldots i_n} c_1^{i_1} \ldots c_n^{i_n},$$

where $r_{i_1 \ldots i_n} \in R$.

Theorem 11.1.14 *Let R be a subring of a commutative ring S such that R and S have the same identity. Let $c \in S$. Then there exists a unique homomorphism α of $R[x]$ onto $R[c]$ such that $\alpha(x) = c$ and $\alpha(a) = a$ for all $a \in R$.*

Proof. Define $\alpha : R[x] \to R[c]$ by $\alpha(\sum a_i x^i) = \sum a_i c^i$ for all $\sum a_i x^i \in R[x]$. Now $a_0 + a_1 x + \cdots + a_n x^n = b_0 + b_1 x + \cdots + b_m x^m$ implies that $n = m$ and $a_i = b_i$ for $i = 1, 2, \ldots, n$. Thus, $a_0 + a_1 c + \cdots + a_n c^n = b_0 + b_1 c + \cdots + b_n c^n$, so α is well defined. By Definition 11.1.13, α clearly maps $R[x]$ onto $R[c]$. Because for any two polynomials $f(x)$, $g(x) \in R[x]$, $k(x) = f(x) + g(x)$ implies $k(c) = f(c) + g(c)$ and $h(x) = f(x)g(x)$ implies $h(c) = f(c)g(c)$, it follows that α preserves $+$ and $\cdot$. Therefore, α is a homomorphism of $R[x]$ onto $R[c]$. Clearly $\alpha(x) = c$ and $\alpha(a) = a$ for all $a \in R$. Let β be a homomorphism of $R[x]$ onto $R[c]$ such that $\beta(x) = c$ and $\beta(a) = a$ for all $a \in R$. Then $\beta(\sum a_i x^i) = \sum \beta(a_i)\beta(x)^i = \sum a_i c^i = \alpha(\sum a_i x^i)$. Thus, $\beta = \alpha$, so α is unique. ■

We emphasize that α is well defined in Theorem 11.1.14 because x is algebraically independent over R. We illustrate this in the following example.

Example 11.1.15 *Define $\alpha : \mathbb{Q}[\sqrt{2}] \to \mathbb{Q}[x]$ by $\alpha(\sum a_i \sqrt{2}) = \sum a_i x^i$. Then α is not a function because $\alpha(2) = 2$ and $\alpha(2) = \alpha((\sqrt{2})^2) = x^2$, but $2 \neq x^2$.*

Worked-Out Exercises

♢ **Exercise 1** Let R be a ring with 1. Show that

$$R[x]/\langle x \rangle \simeq R.$$

Solution: Define $f : R[x] \to R$ by

$$f(a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n) = a_0$$

for all $a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \in R[x]$. Suppose that $a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = b_0 + b_1 x + b_2 x^2 + \cdots + b_m x^m$. Then $a_0 = b_0$, so $f(a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n) = f(b_0 + b_1 x + b_2 x^2 + \cdots + b_m x^m)$. Thus, f is well defined. Clearly f is an epimorphism. Now $a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \in \text{Ker } f$ if and only if $f(a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n) = 0$ if and only if $a_0 = 0$ if and only if $a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \in \langle x \rangle$. Therefore, Ker $f = \langle x \rangle$. Thus,

$$R[x]/\langle x \rangle \simeq R.$$

Exercise 2 Let F be a field and $\alpha : F[x] \to F[x]$ be an automorphism such that $\alpha(a) = a$ for all $a \in F$. Show that $\alpha(x) = ax + b$ for some $a, b \in F$.

Solution: By the division algorithm, $\alpha(x) = g(x)x + b$ for some $g(x) \in F[x]$ and $b \in F$. Because α is onto $F[x]$, there exist $h(x), p(x) \in F[x]$ such that $g(x) = \alpha(h(x))$ and $x = \alpha(p(x))$. Therefore, $\alpha(x) = g(x)x + b = \alpha(h(x))\alpha(p(x)) + \alpha(b) = \alpha(h(x)p(x) + b)$. Thus, $x = h(x)p(x) + b$ because α is one-one. Now $\deg(x) = \deg(h(x)p(x) + b)$ implies that $\deg(h(x)p(x)) = 1$. Hence, either $\deg h(x) = 1$ and $\deg p(x) = 0$ or $\deg h(x) = 0$ and $\deg p(x) = 1$. Suppose $\deg p(x) = 0$. Then $p(x) = c$ for some $c \in F$. This implies that $x = \alpha(p(x)) = \alpha(c) = c$, which is a contradiction. Therefore, $\deg h(x) = 0$ and $\deg p(x) = 1$. Let $h(x) = a$ for some $a \in F$. Thus, $\alpha(x) = \alpha(h(x))x + b = \alpha(a)x + b = ax + b$.

♢ **Exercise 3** Let R be a commutative ring with 1 and $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \in R[x]$. If a_0 is a unit and $a_1, a_2, \ldots, a_n$ are nilpotent elements, prove that $f(x)$ is invertible.

Solution: We prove this result by induction on $n = \deg f(x)$. If $n = 0$, then $f(x) = a_0$. Hence, $f(x)$ is invertible. Assume that the result is true for all polynomials of the above form and degree $< n$. Suppose now $f(x) = a_0+a_1x+a_2x^2+\cdots+a_nx^n \in R[x]$ such that a_0 is a unit and $a_1, a_2, \ldots, a_n$ are nilpotent elements and $\deg f(x) = n$. Let $g(x) = a_0+a_1x+a_2x^2+\cdots+a_{n-1}x^{n-1}$. Note that $\deg g(x) < n$. Hence, by the induction hypothesis, $g(x)$ is invertible. Because a_n is nilpotent there exists a positive integer m such that $a_n^m = 0$. Then $(g(x) + a_nx^n)(g(x)^{-1} - a_ng(x)^{-2}x^n + a_n^2g(x)^{-3}x^{2n} - \cdots + (-1)^{m-1}a_n^{m-1}g(x)^{-(m-1)}x^{(m-1)n}) = 1$. It now follows that $f(x)$ is invertible.

Exercises

1. If I is an ideal of a ring R, prove that $I[x]$ is an ideal of the polynomial ring $R[x]$.
2. Let R be an integral domain. Prove that R and $R[x]$ have the same characteristic.
3. Let R be a commutative ring with 1. Describe, $\langle x \rangle$, the ideal of $R[x]$ generated by x.
4. (i) Let $f(x) = x^4 + 3x^3 + 2x^2 + 2$ and $g(x) = x^2 + 2x + 1 \in \mathbb{Q}[x]$. Find the unique polynomials $q(x)$, $r(x) \in \mathbb{Q}[x]$ such that $f(x) = q(x)g(x) + r(x)$, where either $r(x) = 0$ or $0 \leq \deg r(x) < \deg g(x)$.
   (ii) Let $f(x) = x^4 + [3]x^3 + [2]x^2 + [2]$ and $g(x) = x^2 + [2]x + [1] \in \mathbb{Z}_5[x]$. Find $q(x)$, $r(x) \in \mathbb{Z}_5[x]$ such that $f(x) = q(x)g(x) + r(x)$, where either $r(x) = 0$ or $0 \leq \deg r(x) < \deg g(x)$.
5. Let $f(x) = x^5 + x^4 + x^3 + x + [3]$, $g(x) = x^4 + x^3 + [2]x^2 + [2]x \in \mathbb{Z}_5[x]$. Find $q(x), r(x) \in \mathbb{Z}_5[x]$ such that $f(x) = q(x)g(x) + r(x)$, where either $r(x) = 0$ or $0 \leq \deg\ r(x) < \deg\ g(x)$.
6. Let $R = \mathbb{Z} \oplus \mathbb{Z}$. Show that the polynomial $(1, 0)x$ in $R[x]$ has infinitely many roots in R.
7. Show that the polynomial ring $\mathbb{Z}_4[x]$ over the ring $\mathbb{Z}_4$ is infinite, but $\mathbb{Z}_4[x]$ is of finite characteristic.
8. In the ring $\mathbb{Z}_8[x]$, show that $[1] + [2]x$ is a unit.
9. Let R be a commutative ring with 1 and $f(x) = a_0 + a_1x + \cdots + a_nx^n \in R[x]$. If $f(x)$ is a unit in $R[x]$, prove that a_0 is a unit in R and a_i is nilpotent for all $i = 1, 2, \ldots, n$.
10. Use the result of Exercise 9 to show that $1 + 5x$ is not a unit in $\mathbb{Z}[x]$.
11. Find all units of $\mathbb{Z}[x]$.
12. Find all units of $\mathbb{Z}_6[x]$.
13. Let R be an integral domain. Prove that the units of $R[x]$ are contained in R.
14. In $\mathbb{Z}_8[x]$, prove the following.
    (i) $[4]x^2 + [2]x + [4]$ is a zero divisor.
    (ii) $[2]x$ is nilpotent.
    (iii) $[4]x + [1]$ and $[4]x + [3]$ are units.
15. Let R be a subring of a commutative ring S such that R has an identity.
    (i) In the polynomial ring $R[x_1, x_2, \ldots, x_n]$, prove that $x_1, x_2, \ldots, x_n$ are algebraically independent over R.
    (ii) Prove that the mapping
    $$\alpha : R[x_1, x_2, \ldots, x_n] \to R[c_1, c_2, \ldots, c_n]$$
    defined by $\alpha(\sum_{i_n \ldots i_1} r_{i_1 \ldots i_n} x_1^{i_1} \ldots x_n^{i_n}) = \sum_{i_n \ldots i_1} r_{i_1 \ldots i_n}\ c_1^{i_1} \ldots c_n^{i_n}$ is a homomorphism of $R[x_1, \ldots, x_n]$ onto $R[c_1, \ldots, c_n]$, where $c_1, \ldots, c_n \in S$.
    (iii) Prove that the homomorphism α in (ii) is an isomorphism if and only if $c_1, c_2, \ldots, c_n$ are algebraically independent over R.
16. Let $f(x)$ be a polynomial of degree $n > 0$ in a polynomial ring $K[x]$ over a field K. Prove that any element of the quotient ring $K[x]/\langle f(x) \rangle$ is of the form $g(x) + \langle f(x) \rangle$, where $g(x)$ is a polynomial of degree at most $n - 1$.

17. For the following statements, write the proof if the statement is true; otherwise, give a counterexample.
    (i) If a polynomial ring $R[x]$ has zero divisors, so does R.
    (ii) If R is a field, then $R[x]$ is a field.
    (iii) In $\mathbb{Z}_7[x]$, $(x + [1])^7 = x^7 + [1]$.

Chapter 12

Euclidean Domains

We have seen that both rings $\mathbb{Z}$ and $F[x]$, F a field, have a Euclidean or division algorithm. Because of the significance of these rings and the power of this common property, the concept of a division algorithm is worth abstracting.

12.1 Euclidean Domains

Definition 12.1.1 *A **Euclidean domain** $(E, +, \cdot, v)$ is an integral domain $(E, +, \cdot)$ together with a function $v : E\backslash\{0\} \to \mathbb{Z}^{\#}$ such that*

(i) for all a, $b \in E$ with $b \neq 0$, there exist q, $r \in E$ such that $a = qb + r$, where either $r = 0$ or $v(r) < v(b)$ and

(ii) for all a, $b \in E\backslash\{0\}$, $v(a) \leq v(ab)$.

*v is called a **Euclidean valuation**.*

The next two results show that the ring $\mathbb{Z}$ and the polynomial ring $F[x]$, F a field, are Euclidean domains.

Example 12.1.2 *The ring $\mathbb{Z}$ of integers can be considered a Euclidean domain with $v(a) = |a|$, $a \neq 0$.*

Theorem 12.1.3 *If F is a field, then the polynomial ring $F[x]$ is a Euclidean domain.*

Proof. By Theorem 11.1.2(ii), $F[x]$ is an integral domain. Define

$$v : F[x]\backslash\{0\} \longrightarrow \mathbb{Z}^{\#}$$

by

$$v(f(x)) = \deg f(x)$$

for all $f(x) \in F[x]\backslash\{0\}$. Since $\deg f(x) \geq 0$, $v(f(x)) \in \mathbb{Z}^{\#}$ for all $f(x) \in F[x]\backslash\{0\}$. Let $f(x), g(x) \in F[x]$, $g(x) \neq 0$. By Theorem 11.1.6, there exist $q(x), r(x) \in F[x]$ such that

$$f(x) = q(x)g(x) + r(x), \text{ where either } r(x) = 0 \text{ or } \deg r(x) < \deg g(x).$$

Hence,

$$f(x) \quad = \quad q(x)g(x) + r(x), \text{ where either } r(x) = 0 \text{ or } v(r(x)) < v(g(x)).$$

Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$, $a_n \neq 0$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$, $b_m \neq 0$. Then $f(x)g(x) = a_0b_0 + (a_0b_1 + a_1b_0)x + \cdots + a_nb_mx^{n+m}$. Since F is a field and $a_n \neq 0$, $b_m \neq 0$, we find that $a_nb_m \neq 0$. This implies that $\deg(f(x)g(x)) = n + m$. Thus, $v(f(x)) = \deg(f(x)) = n \leq n + m = \deg(f(x)g(x)) = v(f(x)g(x))$. Hence, $F[x]$ is a Euclidean domain. ■

Example 12.1.4 *Any field can be considered as a Euclidean domain with $v(a) = 1$ for all $a \neq 0$.* ($a = (ab^{-1})b + 0$.)

Definition 12.1.5 *The subset $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$ of the complex numbers is called the set of **Gaussian integers**.*

In the next theorem, we show that $\mathbb{Z}[i]$ is a subring of $\mathbb{C}$ and determine the units of $\mathbb{Z}[i]$. Gauss was the first to study $\mathbb{Z}[i]$ and hence in his honor $\mathbb{Z}[i]$ is called the **ring of Gaussian integers**.

Theorem 12.1.6 *The set $\mathbb{Z}[i]$ of Gaussian integers is a subring of $\mathbb{C}$. The units of $\mathbb{Z}[i]$ are ± 1 and $\pm i$.*

Proof. It is easily verified that $\mathbb{Z}[i]$ is a subring of $\mathbb{C}$. Since $\mathbb{C}$ is a field, $\mathbb{Z}[i]$ is of course an integral domain. Suppose $a+bi$ is a unit of $\mathbb{Z}[i]$. Then there exists $c+di \in \mathbb{Z}[i]$ such that $(a+bi)(c+di) = 1$. This implies that $1 = \overline{1} = \overline{(a+bi)(c+di)} = \overline{(a+bi)}\ \overline{(c+di)} = (a-bi)(c-di)$, where the bar denotes complex conjugate. Thus, $1 = (a^2+b^2)(c^2+d^2)$ and therefore $1 = a^2+b^2$. Hence, $a = 0, b = \pm 1$, or $a = \pm 1, b = 0$, proving that the only units of $\mathbb{Z}[i]$ are ± 1, $\pm i$. ■

Theorem 12.1.7 *The ring $\mathbb{Z}[i]$ of Gaussian integers becomes a Euclidean domain when we let the function,*

$$N : \mathbb{Z}[i]\backslash\{0\} \to \mathbb{Z}^{\#}$$

defined by $N(a+bi) = (a+bi)(a-bi) = a^2+b^2$ for all $a, b \in \mathbb{Z}$, serve as the function v.

Proof. Clearly $N(a+bi)$ is a positive integer for any nonzero element $a+bi \in \mathbb{Z}[i]$. Let $a+bi$, $c+di \in \mathbb{Z}[i]\backslash\{0\}$. Now $N((a+bi)(c+di)) = N(ac-bd+(bc+ad)i) = (ac-bd)^2+(bc+ad)^2 = (a^2+b^2)(c^2+d^2) = N(a+bi)N(c+di)$. From this, it follows that $N(a+bi) \leq N((a+bi)(c+di))$.

It remains to be shown that for $a+bi$ and $c+di \neq 0$ in $\mathbb{Z}[i]$, there exist $q_0+q_1i, r_0+r_1i \in \mathbb{Z}[i]$ such that

$$a+bi = (q_0+q_1i)(c+di) + (r_0+r_1i),$$

where $r_0+r_1i = 0$ or $N(r_0+r_1i) < N(c+di)$. We work backward in order to see how to choose q_0+q_1i. If such an element q_0+q_1i exists, then in $\mathbb{C}$

$$\begin{aligned} r_0+r_1i &= (a+bi) - (c+di)(q_0+q_1i) \\ &= (c+di)[(a+bi)(c+di)^{-1} - (q_0+q_1i)]. \end{aligned}$$

Let $(a+bi)(c+di)^{-1} = u+vi$, where u and v are rational numbers. Then

$$\begin{aligned} r_0+r_1i &= (c+di)[(u+vi) - (q_0+q_1i)] \\ &= (c+di)[(u-q_0) + (v-q_1)i] \\ &= [c(u-q_0) - d(v-q_1)] + [c(v-q_1) + d(u-q_0)]i. \end{aligned}$$

Now

$$\begin{aligned} N(r_0+r_1i) &= [c(u-q_0) - d(v-q_1)]^2 + [c(v-q_1) + d(u-q_0)]^2 \\ &= (c^2+d^2)[(u-q_0)^2 + (v-q_1)^2]. \end{aligned}$$

Hence, $N(r_0+r_1i) < N(c+di)$ if $(u-q_0)^2+(v-q_1)^2 < 1$. We now find an element $q_0+q_1i \in \mathbb{Z}[i]$ so that the latter inequality holds. Take integers q_0 and q_1 such that $(u-q_0)^2 \leq \frac{1}{4}$ and $(v-q_1)^2 \leq \frac{1}{4}$. Then $(u-q_0)^2+(v-q_1)^2 < 1$. Let

$$r_0+r_1i = (a+bi) - (c+di)(q_0+q_1i).$$

Then $a+bi = (c+di)(q_0+q_1i) + (r_0+r_1i)$, where $r_0+r_1i = 0$ or $N(r_0+r_1i) < N(c+di)$. ■

We now consider the ideals of a Euclidean domain.

Recall that an ideal I of a ring R is called a principal ideal if $I = \langle a \rangle$ for some $a \in I$.

Definition 12.1.8 *Let R be a commutative ring with 1. If every ideal of R is a principal ideal, then R is called a* ***principal ideal ring****. An integral domain which is also a principal ideal ring is called a* ***principal ideal domain (PID).***

Theorem 12.1.9 *Every Euclidean domain is a principal ideal domain.*

Proof. Let E be a Euclidean domain with Euclidean valuation v. We want to show that every ideal of E is a principal ideal. Let I be an ideal of E. Since E is a commutative ring with 1, it is enough to show that $I = Ea$ for some $a \in E$. If I is the zero ideal, then $I = E0$. Suppose now $I \neq \{0\}$. Then I contains some nonzero element. Let $P = \{v(x) \mid 0 \neq x \in I\}$. This is a nonempty subset of the nonnegative integers. By the well-ordering principle, we find that P contains a least element. Therefore, there exists an element $a \in I$, $a \neq 0$ such that $v(a) \geq 0$ and $v(a) \leq v(b)$ for all $b \in I$, $b \neq 0$. We now show that $I = Ea$. Since I is an ideal and $a \in I$, it follows that $Ea \subseteq I$. Let $b \in I$. Since E is a Euclidean domain, there exist $q, r \in E$ such that $b = aq+r$, where $r = 0$ or $v(r) < v(a)$. Now $r = b - qa \in I$. If $r \neq 0$, then $v(r) \in P$. This is a contradiction of the minimality of $v(a)$ since $v(r) < v(a)$. Therefore, $r = 0$ and so $b = qa \in Ea$. This proves that $I \subseteq Ea$. Hence, $I = Ea$. ■■

By Theorem 12.1.9, $\mathbb{Z}$, $F[x]$ (F a field), and $\mathbb{Z}[i]$ are principal ideal domains.

Theorem 12.1.10 *Let R be a commutative ring with 1. The following conditions are equivalent.*
(i) R is a field.
(ii) $R[x]$ is a Euclidean domain.
(iii) $R[x]$ is a PID.

Proof. (i)⇒(ii) Follows from Theorem 12.1.3.
(ii)⇒(iii) Follows from Theorem 12.1.9.
(iii)⇒(i) Let $a \in R$ and $a \neq 0$. Consider $I = \langle a, x \rangle$, the ideal of $R[x]$ generated by a and x. Since $R[x]$ is a PID, there exists $f(x) \in R[x]$ such that $I = \langle f(x) \rangle$. Now $a, x \in \langle f(x) \rangle$. Therefore, there exist $g(x)$ and $h(x)$ in $R[x]$ such that $f(x)g(x) = a$ and $f(x)h(x) = x$. Since $f(x)g(x) = a$, we must have $\deg f(x) = 0$ and so $f(x) \in R$. Let $f(x) = b$. Now $bh(x) = x$ implies that $bc = 1$ for some $c \in R$. Thus, b is a unit and so $I = \langle b \rangle = R[x]$. From this, we have $1 \in I$. Therefore, $1 = af_1(x) + xf_2(x)$ for some $f_1(x), f_2(x) \in R[x]$. This implies that $1 = da$ for some $d \in R$. Hence, a is a unit in R and so R is a field. ■

Corollary 12.1.11 *$\mathbb{Z}[x]$ is not a PID.*

Proof. Now $\mathbb{Z}$ is a commutative ring with 1. Since $\mathbb{Z}$ is not a field, $\mathbb{Z}[x]$ is not a PID by Theorem 12.1.10. ■
We conclude this section with the following remark.

Remark 12.1.12 *Consider $\mathbb{Z}[\sqrt{-19}] = \{a + b\sqrt{-19} \mid a, b \in \mathbb{Z}$ and a and b are either both even or both odd$\}$. It is known that $\mathbb{Z}[\sqrt{-19}]$ is a principal ideal domain, but not a Euclidean domain. The proof of this result is beyond the scope of this book. However, the interested reader can find the proof in, J.C. Wilson, "A principal ideal ring that is not a Euclidean ring," Mathematics Magazine* 46(1973), $34 - 38$.

Worked-Out Exercises

♢ **Exercise 1** Let $(E, +, \cdot, v)$ be a Euclidean domain.

(a) Show that $v(a) = v(-a)$ for all $a \in E\backslash\{0\}$.

(b) Show that for all $a \in E\backslash\{0\}$, $v(a) \geq v(1)$, where equality holds if and only if a is a unit in E.

(c) Let n be an integer such that $v(1) + n \geq 0$. Show that the function

$$v_n : E\backslash\{0\} \to \mathbb{Z}^{\#}$$

defined by $v_n(a) = v(a) + n$ for all $a \in E\backslash\{0\}$ is a Euclidean valuation.

Solution: (a) For all $a \in E\backslash\{0\}$, $v(a) = v((-1)(-a)) \geq v(-a) = v((-1)a) \geq v(a)$. Hence, $v(a) = v(-a)$ for all $a \in E\backslash\{0\}$.

(b) Let $a \in E\backslash\{0\}$. Now $v(a) = v(1a) \geq v(1)$. Suppose a is a unit. Then there exists an element $c \in E$ such that $ac = 1$. Thus, $v(1) = v(ac) \geq v(a)$. This implies that $v(a) = v(1)$. Conversely, suppose that $v(a) = v(1)$. Since $a \neq 0$, there exist $q, r \in E$ such that $1 = qa + r$, where $r = 0$ or $v(r) < v(1)$. Now $v(r) < v(1)$ is impossible. Hence, $r = 0$, showing that $1 = qa$. Thus, a is a unit.

(c) Let $a \in E\backslash\{0\}$. Then $v_n(a) = v(a) + n \geq v(1) + n \geq 0$. Hence, $v_n(a) \in \mathbb{Z}^{\#}$. Suppose $a, b \in E$ with $b \neq 0$. There exist $q, r \in E$ such that $a = qb + r$, where either $r = 0$ or $v(r) < v(b)$. Now $v(r) < v(b)$ implies that $v(r) + n < v(b) + n$. Thus, $v_n(r) < v_n(b)$. Also, for $a, b \in E\backslash\{0\}$, $v_n(ab) = v(ab) + n \geq v(a) + n = v_n(a)$. Therefore, v_n is a Euclidean valuation on E.

♢ **Exercise 2** Let n be a square free integer (an integer different from 0 and 1, which is not divisible by the square of any integer). Let $\mathbb{Z}[\sqrt{n}] = \{a + b\sqrt{n} \mid a, b \in \mathbb{Z}\}$. Show that $\mathbb{Z}[\sqrt{n}]$ is an integral domain. Define a function $N : \mathbb{Z}[\sqrt{n}] \to \mathbb{Z}^{\#}$ by

$$N(a + b\sqrt{n}) = (a + b\sqrt{n})(a - b\sqrt{n}) = a^2 - nb^2.$$

(a) Let $x \in \mathbb{Z}[\sqrt{n}]$. Prove that $N(x) = 0$ if and only if $x = 0$.

(b) Prove that $N(xy) = N(x)N(y)$ for all $x, y \in \mathbb{Z}[\sqrt{n}]$.

(c) Let $x \in \mathbb{Z}[\sqrt{n}]$. Prove that $N(x) = \pm 1$ if and only if x is a unit in $\mathbb{Z}[\sqrt{n}]$.

Solution: Let $x = a + b\sqrt{n}$ and $y = c + d\sqrt{n}$ be two elements in $\mathbb{Z}[\sqrt{n}]$. Now $x - y = (a - c) + (b - d)\sqrt{n} \in \mathbb{Z}[\sqrt{n}]$ and $xy = (ac + nbd) + (ad + bc)\sqrt{n} \in \mathbb{Z}[\sqrt{n}]$. We have $0 = 0 + 0\sqrt{n} \in \mathbb{Z}[\sqrt{n}]$ and $1 = 1 + 0\sqrt{n} \in \mathbb{Z}[\sqrt{n}]$. Now it is easy to verify that $\mathbb{Z}[\sqrt{n}]$ is an integral domain.

(a) Let $x = a + b\sqrt{n}$. Then $N(x) = a^2 - nb^2$. Suppose $N(x) = 0$. If $b = 0$, then $a = 0$. If $b \neq 0$, then $n = \frac{a^2}{b^2} = (\frac{a}{b})^2$, which is a contradiction to the assumption that n is a square free integer. Therefore, $a = 0$ and $b = 0$. Thus, $x = 0$. The converse is trivial.

(b) Let $x = a+b\sqrt{n}$ and $y = c+d\sqrt{n}$. Now

$$\begin{aligned} N(xy) &= [(ac+nbd)+(ad+bc)\sqrt{n}][(ac+nbd)-(ad+bc)\sqrt{n}] \\ &= (ac+nbd)^2-(ad+bc)^2n \\ &= a^2c^2+n^2b^2d^2-a^2d^2n-b^2c^2n \\ &= (a^2-nb^2)(c^2-nd^2) \\ &= N(x)N(y). \end{aligned}$$

(c) Let $x = a + b\sqrt{n}$. $N(x) = \pm 1$ if and only if $(a + b\sqrt{n})(a - b\sqrt{n}) = \pm 1$ if and only if $a + b\sqrt{n}$ divides 1, i.e., if and only if $a + b\sqrt{n}$ is a unit in $\mathbb{Z}[\sqrt{n}]$.

♢ **Exercise 3** Show that $\mathbb{Z}[\sqrt{n}]$ is a Euclidean domain for $n = -1, -2, 2, 3$.

Solution: By Worked-Out Exercise 2 (page 187), $\mathbb{Z}[\sqrt{n}]$ is an integral domain. Define $v : \mathbb{Z}[\sqrt{n}]\backslash\{0\} \to \mathbb{Z}^{\#}$ by $v(a + b\sqrt{n}) = |N(a + b\sqrt{n})|$, where N is defined as in Worked-Out Exercise 2. Let $a + b\sqrt{n}$, $c + d\sqrt{n} \in \mathbb{Z}[\sqrt{n}]\backslash\{0\}$. Now

$$\begin{aligned} v((a+b\sqrt{n})(c+d\sqrt{n})) &= |N((a+b\sqrt{n})(c+d\sqrt{n}))| \\ &= \left|(a^2-nb^2)(c^2-nd^2)\right| \\ &= \left|(a^2-nb^2)\right|\left|(c^2-nd^2)\right| \\ &\geq \left|(a^2-nb^2)\right| \\ &= v((a+b\sqrt{n})). \end{aligned}$$

Let $a + b\sqrt{n}$, $c + d\sqrt{n} \in \mathbb{Z}[\sqrt{n}]$ with $c + d\sqrt{n} \neq 0$. We want to show that there exist $q_0 + q_1\sqrt{n}$, $r_0 + r_1\sqrt{n} \in \mathbb{Z}[\sqrt{n}]$ such that

$$a + b\sqrt{n} = (c + d\sqrt{n})(q_0 + q_1\sqrt{n}) + (r_0 + r_1\sqrt{n}),$$

where either $r_0 + r_1\sqrt{n} = 0$ or $\left|(r_0^2 - nr_1^2)\right| < \left|(c^2 - nd^2)\right|$. We work backward in order to see how to choose $q_0 + q_1\sqrt{n}$. If such an element $q_0 + q_1\sqrt{n}$ exists in $\mathbb{Z}[\sqrt{n}]$, then in $\mathbb{Q}[\sqrt{n}]$

$$\begin{aligned} r_0 + r_1\sqrt{n} &= (a + b\sqrt{n}) - (c + d\sqrt{n})(q_0 + q_1\sqrt{n}) \\ &= (c + d\sqrt{n})[(a + b\sqrt{n})(c + d\sqrt{n})^{-1} - (q_0 + q_1\sqrt{n})]. \end{aligned}$$

Let $(a + b\sqrt{n})(c + d\sqrt{n})^{-1} = u + v\sqrt{n}$, where u and v are rational numbers. Then

$$\begin{aligned} r_0 + r_1\sqrt{n} &= (c + d\sqrt{n})[(u + v\sqrt{n}) - (q_0 + q_1\sqrt{n})] \\ &= (c + d\sqrt{n})[(u - q_0) + (v - q_1)\sqrt{n}] \\ &= [c(u - q_0) + d(v - q_1)n] + [c(v - q_1) + d(u - q_0)]\sqrt{n}. \end{aligned}$$

Now

$$\begin{aligned} v(r_0 + r_1\sqrt{n}) &= \left|[c(u - q_0) + d(v - q_1)n]^2 - [c(v - q_1) + d(u - q_0)]^2 n\right| \\ &= \left|(c^2 - nd^2)[(u - q_0)^2 - n(v - q_1)^2]\right| \\ &< \left|(c^2 - nd^2)\right| \end{aligned}$$

if $\left|(u - q_0)^2 - n(v - q_1)^2\right| < 1$. We now find an element $q_0+q_1\sqrt{n} \in \mathbb{Z}[\sqrt{n}]$ such that $\left|(u - q_0)^2 - n(v - q_1)^2\right| < 1$. Take integers q_0 and q_1 such that $(u - q_0)^2 \leq \frac{1}{4}$ and $(v - q_1)^2 \leq \frac{1}{4}$. For $n = -1$ or -2,

$$\left|(u - q_0)^2 - n(v - q_1)^2\right| \leq \frac{1}{4} + (-n)\frac{1}{4} < 1.$$

For $n = 2$ or 3,

$$-\frac{n}{4} \leq (u - q_0)^2 - n(v - q_1)^2 \leq \frac{1}{4}.$$

Then $\left|(u - q_0)^2 - n(v - q_1)^2\right| < 1$ for $n = -1$, -2, 2 or 3. Hence, there exist $q_0 + q_1\sqrt{n}$, $r_0 + r_1\sqrt{n} \in \mathbb{Z}[\sqrt{n}]$ such that

$$a + b\sqrt{n} = (c + d\sqrt{n})(q_0 + q_1\sqrt{n}) + (r_0 + r_1\sqrt{n}),$$

where either $r_0 + r_1\sqrt{n} = 0$ or $\left|(r_0^2 - nr_1^2)\right| < \left|(c^2 - nd^2)\right|$.

♢ **Exercise 4** Let $\mathbb{Z}[i\sqrt{3}] = \{a + bi\sqrt{3} \mid a, b \in \mathbb{Z}\}$. Show that $\mathbb{Z}[i\sqrt{3}]$ is an integral domain. Define $v : \mathbb{Z}[i\sqrt{3}]\backslash\{0\} \to \mathbb{Z}^{\#}$ by $v(a + bi\sqrt{3}) = a^2 + 3b^2$. Show that v is not a Euclidean valuation on $\mathbb{Z}[i\sqrt{3}]$.

Solution: Proceeding as in Worked-Out Exercise 2 (page 187), we can show that $\mathbb{Z}[i\sqrt{3}]$ is an integral domain. Suppose v is a Euclidean valuation. Now 2 and $1 + i\sqrt{3}$ are elements of $\mathbb{Z}[i\sqrt{3}]$. Suppose there exist $q_0 + q_1 i\sqrt{3}$, $r_0 + r_1 i\sqrt{3} \in \mathbb{Z}[i\sqrt{3}]$ such that

$$2 = (1 + i\sqrt{3})(q_0 + q_1 i\sqrt{3}) + (r_0 + r_1 i\sqrt{3}),$$

where either $r_0 + r_1 i\sqrt{3} = 0$ or $r_0^2 + 3r_1^2 < 4$. If $r_0 + r_1 i\sqrt{3} = 0$, then

$$2 = (1 + i\sqrt{3})(q_0 + q_1 i\sqrt{3}).$$

This implies that

$$4 = v(2) = v((1 + i\sqrt{3})(q_0 + q_1 i\sqrt{3})) = 4(q_0^2 + 3q_1^2).$$

Then $q_0^2 + 3q_1^2 = 1$, which shows that $q_0 = \pm 1$, $q_1 = 0$. As a result, $2 = 1 + i\sqrt{3}$ or $2 = -(1 + i\sqrt{3})$, a contradiction. Suppose now $r_0^2 + 3r_1^2 < 4$. Then $r_0^2 + 3r_1^2 = 1$, 2, or 3. Since r_0 and r_1 are integers, $r_0^2 + 3r_1^2 \neq 2$. Suppose $r_0^2 + 3r_1^2 = 1$. Then $r_0 = \pm 1$, $r_1 = 0$. Thus,

$$2 = (1 + i\sqrt{3})(q_0 + q_1 i\sqrt{3}) + (r_0 + r_1 i\sqrt{3}),$$

whence

$$2 = q_0 - 3q_1 + r_0$$

and

$$0 = q_1 + q_0 + r_1.$$

If $r_0 = 1$ and $r_1 = 0$, then $q_0 - 3q_1 = 1$ and $q_1 + q_0 = 0$. This implies that $-2q_1 = 1$, which is impossible. Similarly, for each remaining case we can show a contradiction. Also, from $r_0^2 + 3r_1^2 = 3$, we can show a contradiction. Hence, v is not a Euclidean valuation on $\mathbb{Z}[i\sqrt{3}]$.

Exercises

1. Show that the mapping $v : \mathbb{Z}\backslash\{0\} \to \mathbb{N}$ defined by $v(a) = |a|^n$ for some fixed positive integer n is a Euclidean valuation on $\mathbb{Z}$.
2. In $\mathbb{Z}[\sqrt{3}]$, for $9 + 5\sqrt{3}$ and $1 + 7\sqrt{3}$, find $q_0 + q_1\sqrt{3}$, $r_0 + r_1\sqrt{3} \in \mathbb{Z}[\sqrt{3}]$ such that
$$9 + 5\sqrt{3} = (q_0 + q_1\sqrt{3})(1 + 7\sqrt{3}) + r_0 + r_1\sqrt{3},$$
where either $r_0 + r_1\sqrt{3} = 0$ or $\left|r_0^2 - 3r_1^2\right| < 146$.
3. Consider the integral domain $\mathbb{Z}[i]$. Find $q_0 + q_1 i$, $r_0 + r_1 i \in \mathbb{Z}[i]$ such that
$$3 + 7i = (q_0 + q_1 i)(1 + 2i) + r_0 + r_1 i,$$
where either $r_0 + r_1 i = 0$ or $\left|r_0^2 + r_1^2\right| < 5$.
4. Let $a = 3 + 8i$, $b = -2 + 3i \in \mathbb{Z}[i]$. Find $c, d = x + yi$ in $\mathbb{Z}[i]$ such that $a = bc + d$, where either $d = 0$ or $x^2 + y^2 < 9$.
5. Let $f : R \to S$ be an epimorphism of rings. If R is a principal ideal ring, prove that S is also a principal ideal ring.
6. Prove that the ring $\mathbb{Z}_n$ is a principal ideal ring for all $n \in \mathbb{N}$.
7. Which of the following statements are true? Justify.
   (i) $(\mathbb{Z}, +, \cdot, v)$ is a Euclidean domain, where $v(n) = n^2$ for all $n > 0$.
   (ii) $(\mathbb{Q}, +, \cdot, v)$ is a Euclidean domain, where $v(\frac{p}{q}) = \left|\frac{p}{q}\right|$ for all $\frac{p}{q} \neq 0$.
   (iii) If a ring R is a PID, then every subring of R with identity is a PID.

12.2 Greatest Common Divisors

Definition 12.2.1 *Let R be a commutative ring and a, $b \in R$ be such that $a \neq 0$. If there exists $c \in R$ such that $b = ac$, then a is said to **divide** b or a is said to be a **divisor** of b and we write $a \mid b$.*

When we write $a \mid b$, we mean that $a \neq 0$ and a divides b. The notation $a \nmid b$ will mean that a does not divide b.

Let R be a commutative ring with 1. By Definition 12.2.1, the following results follow immediately. For all $a, b, c \in R$,

(i) $a \mid a$, $1 \mid a$ and $a \mid 0$,

(ii) a is a unit if and only if $a \mid 1$,

(iii) if $a \mid b$ and $b \mid c$, then $a \mid c$.

Definition 12.2.2 *Let R be a commutative ring with 1. A nonzero element $a \in R$ is said to be an **associate** of a nonzero element $b \in R$ if $a = bu$ for some unit $u \in R$.*

Example 12.2.3 *(i) In $\mathbb{Z}$, 1 and -1 are the only units. For every $0 \neq a \in \mathbb{Z}$, a and $-a$ are associates.*

(ii) In $\mathbb{Z}[i]$, $1, -1, i, -i$ are the only units. Thus, $1+i$, $-1-i$, $-1+i$, $1-i$ are all associates of $1+i$.

Example 12.2.4 *In the polynomial ring $F[x]$ over a field F, the units form the set $F\backslash\{0\}$. A nonconstant polynomial $f(x)$ has $uf(x)$ for an associate, where u is a unit in F.*

Theorem 12.2.5 *Let R be a commutative ring with 1 and $a, b, c \in R$.*

(i) If a is an associate of b, then b is an associate of a.

(ii) If a is an associate of b and b is an associate of c, then a is an associate of c.

(iii) Suppose R is an integral domain. Then a is an associate of b if and only if $a \mid b$ and $b \mid a$.

(iv) Suppose R is an integral domain. Then a and b are associates of each other if and only if $\langle a \rangle = \langle b \rangle$.

Proof. (i) This result follows from the fact that the inverse of a unit is also a unit.

(ii) This result follows from the fact that the product of two units is also a unit.

(iii) Suppose a is an associate of b. Then $a = bu$ for some unit $u \in R$. This implies that $b = au^{-1}$. Hence, $a \mid b$ and $b \mid a$. Conversely, suppose that $a \mid b$ and $b \mid a$. Then there exist $q_1, q_2 \in R$ such that $a = q_1 b$ and $b = q_2 a$. Thus, $b = q_2 q_1 b$ and so $1 = q_2 q_1$ by cancellation. This implies that q_1 and q_2 are units and so a and b are associates.

(iv) The result here follows from (iii) and the fact that $\langle a \rangle = \{q_2 a \mid q_2 \in R\}$ and $\langle b \rangle = \{q_1 b \mid q_1 \in R\}$. ■

We now introduce the notion of a greatest common divisor in a commutative ring.

Definition 12.2.6 *Let R be a commutative ring and $a_1, a_2, \ldots, a_n$ be elements in R, not all zero. A nonzero element $d \in R$ is called a **common divisor** of $a_1, a_2, \ldots, a_n$ if $d \mid a_i$ for all $i = 1, 2, \ldots, n$. A nonzero element $d \in R$ is called a **greatest common divisor (gcd)** of $a_1, a_2, \ldots, a_n$ if*

(i) d is a common divisor of $a_1, a_2, \ldots, a_n$ and

(ii) if $c \in R$ is a common divisor of $a_1, a_2, \ldots, a_n$, then $c \mid d$.

The greatest common divisor (gcd) of two elements need not be unique. In fact, the gcd of two elements may not even exist.

Example 12.2.7 *Consider the ring $\mathbb{Z}_{10}$. Then $[4] = [4][6]$ and $[6] = [4][4]$. This shows that $[4]$ and $[6]$ are common divisors of each other. Hence, $[4]$ and $[6]$ must be greatest common divisors of $[4]$ and $[6]$. Now $[4]$ and $[6]$ are associates since $[9]$ is a unit and $[6] = [9][4]$.*

Example 12.2.8 *In the ring $\mathbb{E}$ of even integers, 2 has no divisor. Hence, 2 and no other even integer can have a common divisor.*

Example 12.2.9 *In a field F, $a \mid b$ and $b \mid a$ for all $a, b \in F$ with $a \neq 0$ and $b \neq 0$. Thus, every nonzero element is a gcd of any pair of elements.*

The next result shows that in a principal ideal ring, every pair of elements not both zero has a gcd.

Theorem 12.2.10 *Let R be a principal ideal ring and a, $b \in R$ not both zero. Then a and b have a gcd d. For every gcd d of a and b, there exist s, $t \in R$ such that $d = sa + tb$.*

Proof. The ideal $\langle a, b\rangle$ of R must be a principal ideal, whence there exists $d \in R$ such that $\langle a, b\rangle = \langle d\rangle$. Thus, there exist $u, v \in R$ such that $a = ud$ and $b = vd$. Therefore, d is a common divisor of a and b. Since $d \in \langle a, b\rangle$, there exist $s, t \in R$ such that $d = sa + tb$. Now suppose c is any common divisor of a and b. Then there exist $u', v' \in R$ such that $a = u'c$ and $b = v'c$. Thus, $d = (su' + tv')c$ and so $c \mid d$. Hence, d is a gcd of a and b. Let d' be any gcd of a and b. Then $d \mid d'$ and $d' \mid d$, whence $\langle d'\rangle = \langle d\rangle = \langle a, b\rangle$. Thus, there exist $s', t' \in R$ such that $d' = s'a + t'b$. ■

Corollary 12.2.11 *Let R be a Euclidean domain and $a, b \in R$, not both zero. Then a and b have a gcd d. For every gcd d of a and b, there exist s, $t \in R$ such that $d = sa + tb$.*

Proof. Since every Euclidean domain is a principal ideal ring, the corollary follows by Theorem 12.2.10. ■

Proceeding as in the proof of Theorem 12.2.10, we can prove a similar result for any finite set of elements $a_1, a_2, \ldots, a_n$ (not all zero) of a principal ideal ring.

Let R be an integral domain and $a_1, a_2, \ldots, a_n \in R$, not all zero. Suppose that a gcd of $a_1, a_2, \ldots, a_n$ exists. Let d and d' be two greatest common divisors of $a_1, a_2, \ldots, a_n$. Then $d \mid d'$ and $d' \mid d$. We ask the reader to verify in Exercise 6 (page 193) that d and d' are associates. If d is a gcd of $a_1, a_2, \ldots, a_n$, then any associate of d is also a gcd of $a_1, a_2, \ldots, a_n$. Considering this, we can say that the gcd of $a_1, a_2, \ldots, a_n$ is unique in the sense that if d and d' are greatest common divisors of $a_1, a_2, \ldots, a_n$, then d and d' are associates. Hence, from now on, the gcd of $a_1, a_2, \ldots, a_n$ is denoted by $\gcd(a_1, a_2, \ldots, a_n)$. This outcome motivates the definition of associates. We will further motivate this concept when we examine unique factorization in integral domains.

In a Euclidean domain $(E, +, \cdot, v)$, we have seen that the $\gcd(a, b)$ of two elements $a, b \in E$ (a, b not both zero) exists in E. Next we give an algorithm similar to the algorithm of finding the gcd of two integers given in Chapter 1.

Let $a, b \in E$ with $b \neq 0$.

Step 1: Find q_1 and r_1 in E such that $a = q_1 b + r_1$, where $r_1 = 0$ or $v(r_1) < v(b)$. If $r_1 = 0$, then $b \mid a$ and so $\gcd(a, b) = b$. If $r_1 \neq 0$, then $\gcd(a, b) = \gcd(b, r_1)$. Thus, we need to find $\gcd(b, r_1)$.

Step 2: Find q_2 and r_2 in E such that $b = q_2 r_1 + r_2$, where $r_2 = 0$ or $v(r_2) < v(r_1)$. If $r_2 = 0$, then $\gcd(a, b) = \gcd(b, r_1) = r_1$. If $r_2 \neq 0$, then proceed to find $\gcd(r_1, r_2)$. Since $v(b) > v(r_1) > v(r_2) > \cdots$ is a strictly descending chain of nonnegative integers, the above process must stop after a finite number of steps. Therefore, there exists a positive integer n such that in the nth step there exist elements q_n and r_n in E such that $r_{n-2} = q_n r_{n-1} + r_n$, where $r_n = 0$. Thus,

$$\begin{array}{rcll}
\gcd(a, b) & = & \gcd(b, r_1) & (a = q_1 b + r_1,\ v(r_1) < v(b)) \\
& = & \gcd(r_1, r_2) & (b = q_2 r_1 + r_2,\ v(r_2) < v(r_1)) \\
& = & \gcd(r_2, r_3) & (r_1 = q_3 r_2 + r_3,\ v(r_3) < v(r_2)) \\
& \vdots & \vdots & \vdots \\
& = & \gcd(r_{n-2}, r_{n-1}) & (r_{n-3} = q_{n-1} r_{n-2} + r_{n-1}, \\
& & & \qquad v(r_{n-1}) < v(r_{n-2})) \\
& = & \gcd(r_{n-1}, r_n) & (r_{n-2} = q_n r_{n-1} + r_n,\ r_n = 0).
\end{array}$$

Next we find x, y in E such that $\gcd(a, b) = ax + by$.

$$\begin{array}{rcl}
r_{n-1} & = & r_{n-3} - q_{n-1} r_{n-2} \\
& = & r_{n-3} - q_{n-1}(r_{n-4} - q_{n-2} r_{n-3}) \\
& = & r_{n-3}(1 + (-q_{n-1})(-q_{n-2})) + r_{n-4}(-q_{n-1}) \\
& \vdots & \vdots \\
& = & by + ax.
\end{array}$$

Worked-Out Exercises

♢ **Exercise 1** Let E be a Euclidean domain. Let $a, b, q, r \in E$ be such that $b \neq 0$, $a = qb + r$, and $r \neq 0$. Show that $\gcd(a, b) = \gcd(b, r)$.

Solution: Let $\gcd(a, b) = d$ and $\gcd(b, r) = d'$. Now $d \mid a$ and $d \mid b$. Thus, $r = a - qb$ implies that $d \mid r$. Hence, we find that d is a common divisor of b and r and so $d' \mid d$. Now $d' \mid b$ and $d' \mid r$ and so $a = qb + r$ implies that $d' \mid a$. Therefore, d' is a common divisor of a and b and so $d \mid d'$. By Theorem 12.2.5(iii), it follows that d and d' are associates and so $\gcd(a, b) = \gcd(b, r)$.

Exercise 2 Let a, b, and c be three nonzero elements of a PID R. Show that there exist $x, y \in R$ such that $ax + by = c$ if and only if $\gcd(a, b) \mid c$.

Solution: Let $\gcd(a,b)=d$. Suppose there exist $x,y\in R$ such that $ax+by=c$. Since $d\mid a$ and $d\mid b$, we find that $d\mid c$. Conversely, suppose that $\gcd(a,b)\mid c$. Then $c=dd'$ for some $d'\in R$. Now there exist $x',y'\in R$ such that $d=ax'+by'$. Then $ax'd'+by'd'=dd'=c$. Let $x=x'd'$ and $y=y'd'$. Then $ax+by=c$.

♢ **Exercise 3** In the domain $\mathbb{Z}[i\sqrt{5}]$, prove the following:

(a) $\gcd(2,1+i\sqrt{5})=1$,

(b) gcd of $6(1-i\sqrt{5})$ and $3(1+i\sqrt{5})(1-i\sqrt{5})$ does not exist.

Solution: (a) In $\mathbb{Z}[i\sqrt{5}]$, the units are 1 and -1. Let $a+ib\sqrt{5}=\gcd(2,1+i\sqrt{5})$. Then $(a+ib\sqrt{5})\mid 2$. Thus, $2=(a+ib\sqrt{5})(c+id\sqrt{5})$ for some $c+id\sqrt{5}\in\mathbb{Z}[i\sqrt{5}]$. This implies that

$$4=(a^2+5b^2)(c^2+5d^2).$$

Hence,

$$a^2+5b^2=2,\quad c^2+5d^2=2 \tag{12.1}$$

or

$$a^2+5b^2=4,\quad c^2+5d^2=1 \tag{12.2}$$

or

$$a^2+5b^2=1,\quad c^2+5d^2=4. \tag{12.3}$$

Now Eqs. (12.1) cannot hold for any $c,d\in\mathbb{Z}$. The only integral solutions of $a^2+5b^2=4$ are $a=\pm2$ and $b=0$ and the only integral solutions of $a^2+5b^2=1$ are $a=\pm1$ and $b=0$. Thus, from Eqs. (12.2) and Eqs. (12.3) we find that $\gcd(2,1+i\sqrt{5})=1$ or 2. If $\gcd(2,1+i\sqrt{5})=2$, then $2\mid(1+i\sqrt{5})$. Hence, $1+i\sqrt{5}=2(p+iq\sqrt{5})$ for some $p+iq\sqrt{5}\in\mathbb{Z}[i\sqrt{5}]$. This implies that $2p=1=2q$. But there do not exist integers p and q such that $2p=1=2q$. Therefore, $\gcd(2,1+i\sqrt{5})=1$.

(b) Suppose $\gcd(6(1-i\sqrt{5}),3(1+i\sqrt{5})(1-i\sqrt{5}))$ exists. Then $\gcd(6(1-i\sqrt{5}),3(1+i\sqrt{5})(1-i\sqrt{5}))=3(1-i\sqrt{5})\gcd(2,1+i\sqrt{5})=3(1-i\sqrt{5})$. Now $(1+i\sqrt{5})(1-i\sqrt{5})=6$. Hence, 6 is a common divisor of $6(1-i\sqrt{5})$ and $3(1+i\sqrt{5})(1-i\sqrt{5})$. Consequently, $6\mid 3(1-i\sqrt{5})$. This implies that $2\mid(1-i\sqrt{5})$, which is not true in $\mathbb{Z}[i\sqrt{5}]$. Therefore, $\gcd(6(1-i\sqrt{5}),3(1+i\sqrt{5})(1-i\sqrt{5}))$ does not exist.

♢ **Exercise 4** In $\mathbb{Z}[i]$, find $\gcd(9-5i,-9+13i)$.

Solution: By Theorem 12.1.7, $\mathbb{Z}[i]$ is a Euclidean domain, where the valuation is defined by $N(a+bi)=a^2+b^2$. Now $N(9-5i)=106$ and $N(-9+13i)=250$.

Step 1: $\frac{-9+13i}{9-5i}=\frac{(-9+13i)(9+5i)}{106}=\frac{-81-45i+117i-65}{106}=\frac{-146+72i}{106}=\frac{-146}{106}+\frac{72i}{106}=(-1-\frac{40}{106})+(1-\frac{34}{106})i=(-1+i)-\frac{40+34i}{106}$.

Thus, $-9+13i=(-1+i)(9-5i)-\frac{40+34i}{106}(9-5i)=(-1+i)(9-5i)-\frac{360+306i-200i+170}{106}=(-1+i)(9-5i)-\frac{530+106i}{106}=(-1+i)(9-5i)+(-5-i)$. Note that $N(-5-i)<N(9-5i)$.

Step 2: $\frac{9-5i}{-5-i}=\frac{9-5i}{-5-i}\frac{-5+i}{-5+i}=\frac{-45+9i+25i+5}{26}=\frac{-40+34i}{26}=\frac{-20+17i}{13}=\frac{-20}{13}+\frac{17}{13}i=(-1-\frac{7}{13})+(1+\frac{4}{13})i=(-1+i)+\frac{-7+4i}{13}$.

Thus, $9-5i=(-1+i)(-5-i)+\frac{-7+4i}{13}(-5-i)=(-1+i)(-5-i)+\frac{35+7i-20i+4}{13}=(-1+i)(-5-i)+\frac{39-13i}{13}=(-1+i)(-5-i)+(3-i)$. Note that $N(3-i)<N(-5-i)$.

Step 3: $\frac{-5-i}{3-i}=\frac{-5-i}{3-i}\frac{3+i}{3+i}=\frac{-15-5i-3i+1}{10}=\frac{-14-8i}{10}=\frac{-7-4i}{5}=\frac{-7}{5}-\frac{4i}{5}=(-1-\frac{2}{5})-(1-\frac{1}{5})i=(-1-i)+\frac{-2+i}{5}$.

Thus, $-5-i=(-1-i)(3-i)+\frac{-2+i}{5}(3-i)=(-1-i)(3-i)+\frac{-6+2i+3i+1}{5}=(-1-i)(3-i)+\frac{-5+5i}{5}=(-1-i)(3-i)+(-1+i)$. Note that $N(-1+i)<N(3-i)$.

Step 4: $\frac{3-i}{-1+i}=\frac{3-i}{(-1+i)}\frac{-1-i}{(-1-i)}=\frac{-3-3i+i-1}{2}=\frac{-4-2i}{2}=-2-i$.

Thus, $3-i=(-2-i)(-1+i)+0$.

Hence, $\gcd(9-5i,-9+13i)=-1+i$.

♢ **Exercise 5** In $\mathbb{Z}[x]$, find two polynomials $f(x)$ and $g(x)$ such that $\gcd(f(x),\,g(x))=1$, but there do not exist $f_1(x)$ and $g_1(x)$ in $\mathbb{Z}[x]$ such that $1=f(x)f_1(x)+g(x)g_1(x)$.

Solution: $x+6$ and $x+4$ are elements of $\mathbb{Z}[x]$. The $\gcd(x+6,\,x+4)=1$. Suppose there exist $f_1(x)$ and $g_1(x)$ in $\mathbb{Z}[x]$ such that

$$1=(x+6)f_1(x)+(x+4)g_1(x). \tag{12.4}$$

The constant term of the right-hand side in Eq. (12.4) is an even integer, whereas in the left-hand side, the constant term is 1, a contradiction. Hence, there do not exist $f_1(x)$ and $g_1(x)$ in $\mathbb{Z}[x]$ such that $1=(x+6)f_1(x)+(x+4)g_1(x)$.

♢ **Exercise 6** Let R be a commutative ring with 1 and S denote the set of all infinite sequences $\{a_n\}$ of elements from R. Define $+$ and $\cdot$ on S by

$$\{a_n\} + \{b_n\} = \{a_n + b_n\} \text{ and}$$
$$\{a_n\} \cdot \{b_n\} = \{c_n\},$$

where

$$c_n = a_0b_n + a_1b_{n-1} + \cdots + a_nb_0 \text{ for all } n = 0, 1, 2, \ldots.$$

Show that

(a) S is a commutative ring with 1;

(b) An element $\{a_n\}$ is a unit if and only if a_0 is a unit in R;

(c) If R is a field, then S is a PID.

Solution: (a) It is easy to verify that S is a commutative ring with 1. The sequence $\{1, 0, 0, \ldots\}$ is the identity element of S.

(b) Let $\{a_n\} \in S$. Suppose $\{a_n\}$ is a unit. Then there exists a sequence $\{b_n\}$ such that $\{a_n\}\{b_n\} = 1$. Hence, $a_0b_0 = 1$ and so a_0 is a unit. Conversely, suppose that a_0 is a unit. We now consider the sequence $\{b_n\}$, where $b_0 = a_0^{-1}$, $b_1 = -a_0^{-1}(a_1a_0^{-1})$, $\ldots$, $b_k = -a_0^{-1}(a_1b_{k-1} + \cdots + a_kb_0)$, $k \geq 2$. Now $a_0b_0 = 1$, $a_0b_1 + a_1b_0 = a_0(-a_0^{-1}(a_1a_0^{-1}))+ a_1a_0^{-1} = 0$, $\ldots$, $a_kb_0 + a_{k-1}b_1 + \cdots + a_0b_k = a_kb_0 + a_{k-1}b_1 + \cdots + a_0(-a_0^{-1}(a_1b_{k-1} + \cdots + a_kb_0)) = 0$. Therefore, $\{a_n\}\{b_n\} = 1$, proving that $\{a_n\}$ is a unit.

(c) Suppose R is a field. Let I be an ideal of S. If $I = \{0\}$, then I is a principal ideal. Suppose $I \neq \{0\}$. Let $\{a_n\}$ be a nonzero element of I. We define the order of a nonzero sequence $\{a_n\}$ as the first nonnegative integer n such that $a_n \neq 0$, i.e., n is a nonnegative integer such that $a_n \neq 0$ and $a_i = 0$ for $i < n$. There exists a sequence $\{a_n\}$ such that order of $\{a_n\} \leq$ order of $\{b_n\}$ for all $\{b_n\} \in I$. Suppose order of $\{a_n\} = k$. Let $\{c_n\}$ be a sequence such that $c_i = a_{k+i}$ for all $i \geq 0$. Then $\{c_n\}^{-1}$ exists and $\{c_n\}^{-1}\{a_n\} = \{d_n\} \in I$. Also, $d_k = 1$ and $d_i = 0$ for all $i \neq k$. We now show that $I = \langle\{d_n\}\rangle$. Clearly $\langle\{d_n\}\rangle \subseteq I$. Suppose $\{u_n\} \in I$. Let the order of $\{u_n\}$ be m. Then $m \geq k$. Let $\{r_n\} \in S$ be such that $r_{m-k+i} = u_{m+i}$ for all $i \geq 0$ and $r_i = 0$ for all $i \leq m - k$. It is easy to verify that $\{u_n\} = \{r_n\}\{d_n\} \in \langle\{d_n\}\rangle$. Hence, $I = \langle\{d_n\}\rangle$.

Exercises

1. Find all associates of (i) $3 - 2i$ in $\mathbb{Z}[i]$, (ii) $1 + i\sqrt{5}$ in $\mathbb{Z}[i\sqrt{5}]$, (iii) $[6]$ in $\mathbb{Z}_{10}$, (iv) $[4]$ in $\mathbb{Z}_5$, and (v) $[2] + x$ in $\mathbb{Z}_3[x]$.
2. Find all the units of the integral domain $\mathbb{Z}[i\sqrt{3}]$.
3. Find all the associates of $2 + x - 3x^2$ in $\mathbb{Z}[x]$.
4. Show that $[4]$ and $[6]$ are associates in $\mathbb{Z}_{10}$.
5. Find all units of the polynomial ring $\mathbb{Z}_7[x]$. Find all associates of $x^2 + [2]$ in $\mathbb{Z}_7[x]$.
6. Let R be an integral domain and $a_1, a_2, \ldots, a_n$ $(n \geq 2)$ be elements of R not all zero. If d_1 and d_2 are two greatest common divisors of $a_1, a_2, \ldots, a_n$, prove that d_1 and d_2 are associates.
7. Let $(E, +, \cdot, v)$ be a Euclidean domain. Let $a, b \in E$ be such that a and b are associates. Prove that $v(a) = v(b)$.
8. Let $(E, +, \cdot, v)$ be a Euclidean domain and a, $b \in E$. If $a \mid b$ and $v(a) = v(b)$, prove that a and b are associates.
9. Let $(E, +, \cdot, v)$ be a Euclidean domain and a and b be nonzero elements of E. Prove that $v(ab) > v(a)$ if and only if b is not a unit.
10. Let E be a Euclidean domain. Let a, a', b, b', d be nonzero elements of E such that $a = a'd$ and $b = b'd$. Prove that $\gcd(a', b') = 1$ if and only if $\gcd(a, b) = d$.
11. In a PID R, prove that the congruence $ax \equiv b (\text{mod } c)$, where a, b, c are nonzero elements of R has a solution in R if and only if $\gcd(a, c) \mid b$. (Here $ax \equiv b (\text{mod } c)$ means $ax - b = cr$ for some $r \in R$.)
12. Let R be an integral domain. Let a, b, and c be nonzero elements of R such that $\gcd(a, b)$ and $\gcd(ca, cb)$ exist. Prove that $\gcd(ca, cb) = c\gcd(a, b)$.
13. In $\mathbb{Z}[i]$, find $\gcd(2-7i, 2+11i)$. Also, find x and y in $\mathbb{Z}[i]$ such that $\gcd(2-7i, 2+11i) = x(2-7i)+y(2+11i)$.

14. Let R be an integral domain and $a_1, a_2, \ldots, a_n$ $(n \geq 2)$ be nonzero elements of R. An element $d \in R$ is called a **least common multiple (lcm)** of $a_1, a_2, \ldots, a_n$ if
(i) $a_i \mid d$, $i = 1, 2, \ldots n$ and
(ii) if $c \in R$ is such that $a_i \mid c$, $i = 1, 2, \ldots n$, then $d \mid c$.
Prove the following in R.
(i) If d_1 and d_2 are two least common multiples of $a_1, a_2, \ldots, a_n$, then d_1 and d_2 are associates.
(ii) If d is a least common multiple of $a_1, a_2, \ldots, a_n$, then rd is a least common multiple of $ra_1, ra_2, \ldots, ra_n$, for all $r \in R$, $r \neq 0$.
15. Let I be the set of all nonunits of $\mathbb{Z}[i]$. Is I an ideal of $\mathbb{Z}[i]$? Show that for any nontrivial ideal P of $\mathbb{Z}[i]$, the quotient ring $\mathbb{Z}[i]/P$ is a finite ring.
16. Show that $\mathbb{Z}[\sqrt{2}]$ has no unit between 1 and $1 + \sqrt{2}$.
17. In the domain $\mathbb{Z}[\sqrt{2}]$, prove that an element $a + b\sqrt{2} \neq \pm 1$ is a unit if and only if $a + b\sqrt{2} = (1 + \sqrt{2})^k$ or $a + b\sqrt{2} = -(1 + \sqrt{2})^k$ for some positive integer k.
18. An integral domain R is said to satisfy the **gcd property** if every finite nonempty subset of R has a gcd. Prove that every PID satisfies the gcd property.
19. Prove that the integral domain $\mathbb{Z}[\sqrt{2}]$ satisfies the gcd property, where the gcd property is defined in Exercise 18.

12.3 Prime and Irreducible Elements

In this section, we introduce the concepts of prime elements and irreducible elements in a commutative ring with 1. We show that in a PID and hence in a Euclidean domain these two concepts coincide.

Definition 12.3.1 *Let R be a commutative ring with* 1.
(i) An element p of R is called ***irreducible*** *if p is nonzero and a nonunit, and $p = ab$ with $a, b \in R$ implies that either a or b is a unit. An element p of R is called* ***reducible*** *if p is not irreducible.*
(ii) An element p of R is called ***prime*** *if p is nonzero and a nonunit, and if whenever $p \mid ab$, $a, b \in R$, then either p divides a or p divides b.*
(iii) Two elements a and b of R are called ***relatively prime*** *if their only common divisors are units.*

Remark 12.3.2 *Let $p \in \mathbb{Z}$. If p is an ordinary prime, then both p and $-p$ are irreducible and prime in the sense of Definition* 12.3.1.

From the definition of an irreducible element, it follows that the only divisors of an irreducible element p are the associates of p and the unit elements of R. The converse of this result does not always hold in a commutative ring with 1.

Example 12.3.3 *The ring $\mathbb{Z}_6$ is a commutative ring with* 1. *In this ring, the unit elements are* $[1]$ *and* $[5]$. *Since* $[3] = [3][3]$ *and* $[3]$ *is not a unit it follows that* $[3]$ *is not irreducible. But* $[3]$ *is an associate of* $[3]$. *Also, in $\mathbb{Z}_6$, it can be verified that* $[3]$ *is divisible only by associates and the units of $\mathbb{Z}_6$. Next, we show that* $[3]$ *is a prime element in $\mathbb{Z}_6$. Let $[a], [b] \in \mathbb{Z}_6$ and $[3] \mid [a][b]$. Then there exists $[c] \in \mathbb{Z}_6$ such that $[a][b] = [3][c]$, i.e., $[ab] = [3c]$. From this, it follows that $6 \mid (ab - 3c)$. This implies that $3 \mid (ab - 3c)$. Since $3 \mid 3c$, we must have $3 \mid ab$. Since 3 is prime in $\mathbb{Z}$, $3 \mid a$ or $3 \mid b$. Thus, either $[3] \mid [a]$ or $[3] \mid [b]$. Hence, $[3]$ is a prime element in $\mathbb{Z}_6$.*

Theorem 12.3.4 *Let R be an integral domain and $p \in R$ be such that p is nonzero and a nonunit. Then p is irreducible if and only if the only divisors of p are the associates of p and the unit elements of R.*

Proof. Suppose the only divisors of p are the associates of p and the unit elements of R. Let $p = ab$ for some $a, b \in R$. Suppose a is not a unit. Then a is an associate of p. Therefore, $a = pu$ for some unit $u \in R$. Now $p = pub$. Since R is an integral domain, it follows that $ub = 1$. Hence, b is a unit and so p is irreducible. We leave the converse as an exercise. ■

We now consider several examples of prime elements and irreducible elements.

Example 12.3.5 *In $\mathbb{Z}$,* 1 *and* -1 *are the only units, and therefore* 2 *is divisible by* ± 1 *and* ± 2. *It follows that* 2 *is not divisible by any other integer. Therefore,* 2 *is an irreducible element. Suppose now $2 \mid ab$ and* 2 *does not divide a for some $a, b \in \mathbb{Z}$. Since* 2 *does not divide a, a is an odd integer and so* $\gcd(2, a) = 1$. *Therefore, there exist $c, d \in \mathbb{Z}$ such that $1 = 2c + ad$. Thus, $b = 2cb + abd$. Since $2 \mid ab$ and $2 \mid 2bc$, it follows that $2 \mid b$. Hence,* 2 *is prime.*

Example 12.3.6 *The polynomial x^2+1 is irreducible in $\mathbb{R}[x]$, but is reducible in $\mathbb{C}[x]$. If x^2+1 were reducible in $\mathbb{R}[x]$, then there would exist real numbers a, b, c, d such that*

$$x^2+1=(ax+b)(cx+d)=acx^2+(ad+bc)x+bd.$$

Then $ac=1=bd$ and $ad+bc=0$. Thus, $1=(ac)(bd)=(ad)(bc)=(ad)(-ad)$. Hence, $1=-(ad)^2$, which is impossible in $\mathbb{R}$. However, $x^2+1=(x+i)(x-i)$ in $\mathbb{C}[x]$.

Example 12.3.7 *The polynomial x^2-2 is irreducible in $\mathbb{Q}[x]$ and reducible in $\mathbb{R}[x]$. If x^2-2 were reducible in $\mathbb{Q}[x]$, then there would exist $a,b,c,d\in\mathbb{Q}$ such that*

$$x^2-2=(ax+b)(cx+d)=acx^2+(ad+bc)x+bd.$$

Then $ac=1, ad+bc=0$, and $bd=-2$. Thus, $(ad)^2=(ad)(ad)=-(ad)(bc)=(ac)(-bd)=2$. This implies that $\sqrt{2}=ad\in\mathbb{Q}$. This is a contradiction since $\sqrt{2}\notin\mathbb{Q}$. Therefore, x^2-2 is irreducible in $\mathbb{Q}[x]$. However, $x^2-2=(x-\sqrt{2})(x+\sqrt{2})$ in $\mathbb{R}[x]$.

Example 12.3.8 *The polynomial $ax+b$ is irreducible in $F[x]$, where F is a field and $a\neq 0$. Suppose $ax+b=f(x)g(x)$. Then $\deg(f(x)g(x))=1=\deg f(x)+\deg g(x)$. We may assume that $\deg f(x)=0$ and $\deg g(x)=1$. Since $\deg f(x)=0$, $f(x)$ is a nonzero constant polynomial and thus a unit. Hence, $ax+b$ is irreducible.*

Example 12.3.9 *Consider the polynomial ring $\mathbb{Z}[x,y]$. Then x and y are irreducible. $2x$ is not prime since $2x\mid 2x$, but $2x$ does not divide 2 and $2x$ does not divide x. Also, $2x$ is reducible. x^2 and y^2 are relatively prime, but neither is irreducible nor prime.*

Theorem 12.3.10 *Let R be an integral domain and p be a prime element in R. Then p is irreducible.*

Proof. Suppose $p=bc$ for some $b,c\in R$. To show p is irreducible, we must show that either b is a unit or c is a unit. Now $p=bc$ implies that $p\mid bc$. Since p is prime, $p\mid b$ or $p\mid c$. If $p\mid b$, then $b=pq$ for some $q\in R$. Thus, $p=bc=pqc$ and so $p(1-qc)=0$. Since R is an integral domain and $p\neq 0$, $p(1-qc)=0$ and so $1-qc=0$. Thus, $qc=1$, which implies that c is a unit. Similarly, if $p\mid c$, then b is a unit. Hence, p is irreducible. ■

The following example shows that the converse of Theorem 12.3.10 is not true.

Example 12.3.11 *Consider the integral domain*

$$\mathbb{Z}[i\sqrt{5}]=\{a+bi\sqrt{5}\mid a,b\in\mathbb{Z}\}.$$

Let us show that $3=3+0i\sqrt{5}\in\mathbb{Z}[i\sqrt{5}]$ is irreducible, but not prime. Suppose $3=(a+bi\sqrt{5})(c+di\sqrt{5})$ in $\mathbb{Z}[i\sqrt{5}]$. Then $3=\overline{3}=(a-bi\sqrt{5})(c-di\sqrt{5})$. Hence, $9=(a^2+5b^2)(c^2+5d^2)$. Since a,b,c,d are integers, the previous equality implies that

$$a^2+5b^2=3 \text{ and } c^2+5d^2=3 \tag{12.5}$$

or

$$a^2+5b^2=1 \text{ and } c^2+5d^2=9 \tag{12.6}$$

or

$$a^2+5b^2=9 \text{ and } c^2+5d^2=1. \tag{12.7}$$

Clearly there do not exist integers a,b,c,d satisfying Eqs. (12.5). The first equation of Eqs. (12.6) implies that $b=0$ and $a=\pm 1$. Thus, it follows that $a+bi\sqrt{5}$ is a unit. Similarly, the second equation of Eqs. (12.7) implies that $c+di\sqrt{5}$ is a unit. Hence, 3 is irreducible. Now $3\mid 6$ and $6=(1+i\sqrt{5})(1-i\sqrt{5})$. Suppose $3\mid(1+i\sqrt{5})$. Then $1+i\sqrt{5}=3(a+bi\sqrt{5})$ for some $a,b\in\mathbb{Z}$. This implies that $3a=1$, a contradiction, since the equation $3a=1$ has no solution in $\mathbb{Z}$. Hence, 3 does not divide $(1+i\sqrt{5})$. Similarly, 3 does not divide $(1-i\sqrt{5})$. Thus, 3 is not prime.

The following theorem show that the converse of Theorem 12.3.10 holds in a principal ideal ring.

Theorem 12.3.12 *Let R be a principal ideal ring and $p\in R$. If p is irreducible, then p is prime.*

Proof. Suppose p divides ab, where $a,b\in R$. Then there exists $r\in R$ such that $pr=ab$. Now $\langle p,b\rangle=\langle d\rangle$ for some $d\in R$. Therefore, there exists $q\in R$ such that $p=dq$. Since p is irreducible, either d or q must be a unit. If d is a unit, then $\langle p,b\rangle=\langle d\rangle=R$. Hence, $1=sp+tb$ for some $s,t\in R$. Therefore, $a=asp+atb=asp+tpr=(as+tr)p$. This implies that p divides a. If, on the other hand, q is a unit, then $d=pq^{-1}\in\langle p\rangle$. Thus, $\langle d\rangle\subseteq\langle p\rangle\subseteq\langle p,b\rangle=\langle d\rangle$ so that $\langle p\rangle=\langle p,b\rangle$. Hence, $b\in\langle p\rangle$ and so p divides b. ■

Corollary 12.3.13 *Let R be a principal ideal domain and $p \in R$. Then p is irreducible if and only if p is prime.*

Proof. The result follows by Theorems 12.3.10 and 12.3.12. ■

Corollary 12.3.14 *Let R be a Euclidean domain and $p \in R$. Then p is irreducible if and only if p is prime.*

Proof. Since every Euclidean domain is a principal domain, the result follows from Corollary 12.3.13. ■

Theorem 12.3.15 *Let R be a principal ideal ring and a, $b \in R$. If a and b are relatively prime, then there exist s, $t \in R$ such that $1 = sa + tb$.*

Proof. Since the common divisors are units, 1 is a gcd of a and b. The desired result follows from Theorem 12.2.10. ■

We conclude this section by proving the following theorem, which characterizes irreducible polynomials over a field.

Theorem 12.3.16 *Consider the polynomial ring $F[x]$ over the field F and $p(x) \in F[x]$. Then the following conditions are equivalent.*

(i) $p(x)$ is irreducible.

(ii) $F[x]/\langle p(x)\rangle$ is an integral domain.

(iii) $F[x]/\langle p(x)\rangle$ is a field.

Proof. (i)⇒(iii). Let $\overline{f(x)} \in F[x]/\langle p(x)\rangle$ be such that $\overline{f(x)} \neq \overline{0}$, where $\overline{f(x)}$ denotes the coset $f(x) + \langle p(x)\rangle$. Now $up(x)$ and u, where $u \in F\backslash\{0\}$, are the only elements of $F[x]$ which divide $p(x)$. Since $f(x) \notin \langle p(x)\rangle$, $f(x)$ and $p(x)$ are relatively prime and so there exist $s(x)$, $t(x) \in F[x]$ such that $1 = s(x)f(x) + t(x)p(x)$. Thus

$$\overline{1} = \overline{s(x)f(x)} + \overline{t(x)p(x)} \text{ (in } F[x]/\langle p(x)\rangle)$$

and so $\overline{1} = \overline{s(x)}\ \overline{f(x)}$. Hence, $\overline{f(x)}$ has an inverse, namely, $\overline{s(x)}$, and so $F[x]/\langle p(x)\rangle$ is a field.

(iii)⇒(ii): Immediate.

(ii)⇒(i): If $p(x)$ is a unit, then $\langle p(x)\rangle = F[x]$ and so $F[x]/\langle p(x)\rangle = \{0\}$, a contradiction to the hypothesis that $F[x]/\langle p(x)\rangle$ is an integral domain. Therefore, $p(x)$ is not a unit. Suppose $p(x) = f(x)g(x)$. Then $\overline{0} = \overline{p(x)} = \overline{f(x)g(x)} = \overline{f(x)}\ \overline{g(x)}$. Therefore, $\overline{f(x)} = \overline{0}$ or $\overline{g(x)} = \overline{0}$. This implies that $f(x) \in \langle p(x)\rangle$ or $g(x) \in \langle p(x)\rangle$, say, $f(x) \in \langle p(x)\rangle$. Thus, $f(x) = q(x)p(x)$ for some $q(x) \in F[x]$. Hence, $p(x) = q(x)p(x)g(x)$ and so by a degree argument $q(x)$, $g(x) \in F\backslash\{0\}$ are units. Thus, the only factorization of $p(x)$ is $u^{-1}(up(x))$, where u is a unit in $F[x]$. Consequently, $p(x)$ is irreducible. ■

Worked-Out Exercises

♢ **Exercise 1** Show that $[2]$ is a prime element in $\mathbb{Z}_{10}$, but $[2]$ is not irreducible in $\mathbb{Z}_{10}$.

Solution: In $\mathbb{Z}_{10}$, $[1]$, $[3]$, $[7]$, and $[9]$ are the only units. Now $[2] = [2] \cdot [6]$. Since neither $[2]$ nor $[6]$ is a unit, $[2]$ is reducible. Suppose $[2] \mid [a][b]$. Then $[2] \mid [ab]$. Therefore, $[ab] = [k][2]$ for some $[k] \in \mathbb{Z}_{10}$. This implies that $ab - 2k = 10r$ for some $r \in \mathbb{Z}$, i.e., $ab = 2k + 10r = 2(k + 5r)$. Therefore, $2 \mid ab$. Since 2 is prime in $\mathbb{Z}$, $2 \mid a$ or $2 \mid b$. Hence, $[2] \mid [a]$ or $[2] \mid [b]$. Thus, $[2]$ is prime. Note that $\mathbb{Z}_{10}$ is not an integral domain.

♢ **Exercise 2** Let R be an integral domain such that any two elements $a, b \in R$, not both zero, have a gcd d expressible in the form $d = ra + tb$, $r, t \in R$. Let $p \in R$. Show that p is prime if and only if p is irreducible.

Solution: Every prime element in an integral domain is irreducible by Theorem 12.3.10. Let us prove the converse. Suppose p is irreducible. Let $p \mid ab$, $a, b \in R$. Now $\gcd(p, a)$ exists in R. Let $d = \gcd(p, a)$. Since $d \mid p$ and p is irreducible, it follows that either d is an associate of p or d is a unit. Suppose d is an associate of p. Then $p \mid d$. This implies that $p \mid a$, since $d \mid a$. Suppose d is a unit. Since 1 is an associate of d, $1 = \gcd(p, a)$. Thus, there exist $s, t \in R$ such that $1 = ps + at$. This implies that $b = psb + abt$. Now $p \mid psb$ and $p \mid abt$. Hence, $p \mid b$.

♢ **Exercise 3** Let n be a square free integer (an integer different from 0 and 1, which is not divisible by the square of any integer). Let $\mathbb{Z}[\sqrt{n}] = \{a + b\sqrt{n} \mid a, b \in \mathbb{Z}\}$. Define a function $N : \mathbb{Z}[\sqrt{n}] \to \mathbb{Z}$ by

$$N(a + b\sqrt{n}) = (a + b\sqrt{n})(a - b\sqrt{n}) = a^2 - nb^2.$$

Show that if $N(x)$ is a prime integer, then x is irreducible for all $x \in \mathbb{Z}[\sqrt{n}]$.

Solution: Suppose $N(x) = p$, where p is a prime integer. Suppose $x = (a+b\sqrt{n})(c+d\sqrt{n})$. Now $p = N(a+b\sqrt{n})N(c+d\sqrt{n}) = (a^2 - nb^2)(c^2 - nd^2)$ by Worked-Out Exercise 2 (page 187). Hence, either $(a^2 - nb^2) = \pm 1$ or $(c^2 - nd^2) = \pm 1$, i.e., either $a + b\sqrt{n}$ is a unit or $c + d\sqrt{n}$ is a unit. Thus, x is irreducible.

Exercises

1. Show that in the integral domain $\mathbb{Z}[i\sqrt{5}]$, $2 + i\sqrt{5}$ is an irreducible element, but not a prime element.
2. Show that $2 - i$, $1 + i$, and 11 are irreducible elements in $\mathbb{Z}[i]$.
3. In $\mathbb{Z}[i\sqrt{5}]$, show that 3 is not a prime element.
4. In $\mathbb{Z}_{12}$, show that $[3]$ is a prime element, but is not irreducible.
5. Is the polynomial $x^2 + [1]$ irreducible in $\mathbb{Z}_2[x]$?
6. Let T be the set of all sequences $\{a_n\}$ of elements of $\mathbb{Z}$. Prove the following.
   (i) T is an integral domain with respect to addition and multiplication defined by for all $\{a_n\}, \{b_n\} \in T$,
   $$\begin{aligned} \{a_n\} + \{b_n\} &= \{a_n + b_n\} \\ \{a_n\} \cdot \{b_n\} &= \{c_n\}, \qquad \text{where } c_n = \textstyle\sum_{i=0}^{n} a_i b_{n-i}. \end{aligned}$$
   (ii) $T_0 = \{\{a_n\} \in T \mid a_i = 0 \text{ for all but a finite number of indices}\}$ is a subring with identity.
   (iii) The element $(1, 1, 0, \ldots)$ is a unit in T, but not in T_0.
   (iv) $(2, 3, 1, 0, 0, \ldots)$ is irreducible in T, but not in T_0.
7. Let R be an integral domain. Show that (i) every associate of an irreducible element in R is irreducible and (ii) every associate of a prime element in R is prime.
8. In $\mathbb{Z}[i]$, show that 3 is a prime element, but 5 is not a prime element.
9. What are the prime elements of $\mathbb{Z}_9$? Are they irreducible?
10. In $\mathbb{Z}[i]$, if $a + bi$ is an element such that $a^2 + b^2$ is a prime integer, then show that $a + bi$ is a prime element.
11. Let $a + bi\sqrt{3} \in \mathbb{Z}[i\sqrt{3}]$. If $a^2 + 3b^2$ is a prime integer, show that $a + bi\sqrt{3}$ is an irreducible element in $\mathbb{Z}[i\sqrt{3}]$.
12. In the following exercises, write the proof if the statement is true; otherwise, give a counterexample.
    (i) 13 is an irreducible element in $\mathbb{Z}[i]$.
    (ii) Every prime element of $\mathbb{Z}$ is also a prime element of $\mathbb{Z}[i]$.
    (iii) In $\mathbb{Z}_{18}$, every prime element is an irreducible element.
    (iv) In $\mathbb{Z}[i]$, $a + bi$ is a prime element if and only if $a - bi$ is a prime element.
    (v) In a PID R, if p and q are two prime elements such that $p \mid q$, then p and q are associates.

Chapter 13

Unique Factorization Domains

13.1 Unique Factorization Domains

In this section, we study those integral domains in which an analogue of the fundamental theorem of arithmetic holds.

Definition 13.1.1 *A nonzero nonunit element a of an integral domain D is said to have a* ***factorization*** *if a can be expressed as*

$$a = p_1 p_2 \cdots p_n,$$

where $p_1, p_2, \ldots, p_n$ are irreducible elements of D. The expression $p_1 p_2 \cdots p_n$ is called a ***factorization*** *of a.*

An integral domain D is called a **factorization domain (FD)** if every nonzero nonunit element has a factorization.

In Chapter 15, we saw that in an integral domain D every nonzero element $a \in D$ is always divisible by the associates of a and the units of D. These are called the **trivial factors** of a. All other factors (if any) of a are called **nontrivial**. For example, ± 2 and ± 3 are nontrivial factors of 6 in $\mathbb{Z}$. In the following lemma, we show that a nonzero nonunit element that has no factorization as a product of irreducible elements can be expressed as a product of any number of nontrivial factors.

Lemma 13.1.2 *Let D be an integral domain. Let a be a nonzero nonunit element of D such that a does not have a factorization. Then for every positive integer n, there exist nontrivial factors $a_1, a_2, \ldots, a_n \in D$ of a such that $a = a_1 a_2 \cdots a_n$.*

Proof. By the hypothesis, a is not irreducible. Therefore, $a = a_1 b_1$, where a_1, $b_1 \in D$ are nontrivial factors of a. At least one of a_1 or b_1 does not have a factorization; otherwise the factorization of a_1 and b_1 put together produces a factorization of a. Suppose a_1 does not have a factorization. Then a_1 is a nonzero nonunit element and a_1 is not irreducible. There exist nontrivial factors a_2, $b_2 \in D$ of a_1 such that $a_1 = a_2 b_2$. Then $a = a_2 b_2 b_1$. Now at least one of a_2 or b_2 does not have a factorization. If a_2 does not have a factorization, we repeat the above process with a_2. Proceeding this way, we can find nontrivial factors $a_1, a_2, \ldots, a_n \in D$ of a such that $a = a_1 a_2 \cdots a_n$. ■

Theorem 13.1.3 *Let D be an integral domain with a function $N : D\backslash\{0\} \to \mathbb{Z}^{\#}$ such that for all $a, b \in D\backslash\{0\}$, $N(ab) \geq N(b)$, where equality holds if and only if a is a unit. Then D is a FD.*

Proof. Suppose D contains a nonzero nonunit element a such that a does not have a factorization. Now $N(a) \in \mathbb{Z}^{\#}$. Let $N(a) = n$. By Lemma 13.1.2, a can be expressed as a product of $n + 2$ nontrivial factors $a_1, a_2, \ldots, a_{n+2} \in D$. Then $a = a_1 a_2 \cdots a_{n+2}$ and

$$\begin{aligned} n &= N(a) \\ &> N(a_2 \cdots a_{n+2}) \quad \text{(since } a_1 \text{ is not a unit)} \\ &> N(a_3 \cdots a_{n+2}) \\ &> N(a_4 \cdots a_{n+2}) \\ &\vdots \\ &> N(a_{n+1} a_{n+2}) \\ &> N(a_{n+2}). \end{aligned}$$

This shows that there exist at least $n+1$ distinct nonnegative integers strictly less than n, a contradiction. Thus, D is a FD. ■

Example 13.1.4 *Consider the integral domain $\mathbb{Z}[i]$. Define*

$$N : \mathbb{Z}[i]\backslash\{0\} \to \mathbb{Z}^{\#}$$

by $N(a+bi) = a^2 + b^2$ for all $a+bi \in \mathbb{Z}[i]$. It is easy to verify that $a+bi$ is a unit if and only if $N(a+bi) = 1$. Let $a+bi$, $c+di$ be two nonzero elements of $\mathbb{Z}[i]$. Then $N((a+bi)(c+di)) = N((ac-bd)+(ad+bc)i) = (ac-bd)^2 + (ad+bc)^2 = (a^2+b^2)(c^2+d^2) \geq (c^2+d^2) = N(c+di)$, where the equality holds if and only if $N(a+bi)$ is a unit. Hence, $\mathbb{Z}[i]$ is a FD.

Definition 13.1.5 *An integral domain D is said to satisfy the **ascending chain condition for principal ideals (ACCP),** if for each sequence of principal ideals, $\langle a_1\rangle, \langle a_2\rangle, \langle a_3\rangle, \ldots$ such that*

$$\langle a_1\rangle \subseteq \langle a_2\rangle \subseteq \langle a_3\rangle \subseteq \cdots,$$

there exists a positive integer n (depending on the sequence) such that $\langle a_n\rangle = \langle a_t\rangle$ for all $t \geq n$.

Lemma 13.1.6 *Every principal ideal domain D satisfies the ACCP.*

Proof. Let $\langle a_1\rangle \subseteq \langle a_2\rangle \subseteq \langle a_3\rangle \subseteq \cdots$ be a chain of principal ideals in D. It can be easily verified that $I = \cup_{i\in\mathbb{N}} \langle a_i\rangle$ is an ideal of D. Since D is a PID, there exists an element $a \in D$ such that $I = \langle a\rangle$. Hence, $a \in \langle a_n\rangle$ for some positive integer n. Then $I \subseteq \langle a_n\rangle \subseteq I$. Therefore, $I = \langle a_n\rangle$. For $t \geq n$, $\langle a_t\rangle \subseteq I = \langle a_n\rangle \subseteq \langle a_t\rangle$. Thus, $\langle a_n\rangle = \langle a_t\rangle$ for all $t \geq n$. ■

Theorem 13.1.7 *An integral domain D with the ACCP is a FD.*

Proof. Suppose D is not a FD. Then there exists a nonzero nonunit element a such that a does not have a factorization. Thus, a is not irreducible and so $a = a_1b_1$, where $a_1, b_1 \in D$ are nontrivial factors of a. At least one of a_1 or b_1 must not have a factorization, otherwise the factorization of a_1 and b_1 put together will produce a factorization of a. Suppose a_1 does not have a factorization. Now a and a_1 are not associates. Therefore, $\langle a\rangle \subset \langle a_1\rangle$. Since a_1 does not have a factorization, we can express $a_1 = a_2b_2$, where $a_2, b_2 \in D$ are nontrivial factors of a_1. At least one of a_2 or b_2 does not have a factorization. Suppose a_2 does not have a factorization. Then $\langle a\rangle \subset \langle a_1\rangle \subset \langle a_2\rangle$. We now repeat the above process with a_2. Thus, we find that there exists an infinite strictly ascending chain of principal ideals in D, a contradiction. Hence, D is a FD. ■

Corollary 13.1.8 *Every PID is a FD.*

Proof. The proof is immediate by Lemma 13.1.6 and Theorem 13.1.7. ■

Definition 13.1.9 *An integral domain D is called a **unique factorization domain** (**UFD**) if the following two conditions hold in D :*

(i) every nonzero nonunit element of D can be expressed as

$$a = p_1p_2\cdots p_n,$$

where $p_1, p_2, \ldots, p_n$ are irreducible elements of D and

(ii) if $a = p_1p_2\cdots p_n = q_1q_2\cdots q_m$ are two factorizations of a as a finite product of irreducible elements of D, then $n = m$ and there is a permutation σ of $\{1, 2, \ldots, n\}$ such that p_i and $q_{\sigma(i)}$ are associates for all $i = 1, 2, \ldots, n$.

From the above definition, it follows that an integral domain D is a UFD if and only if D is a FD and every nonzero nonunit element of D is uniquely expressible (apart from unit factors and order of the factors) as a finite product of irreducible elements.

Let us first prove the following interesting property of a UFD.

Theorem 13.1.10 *In a unique factorization domain, every irreducible element is prime.*

Proof. Let D be a UFD. Let p be an irreducible element of D and $p \mid ab$ in D, where $a, b \in D$. If $a = 0$, then p divides a, and if $b = 0$, then p divides b. If a is a unit, then p divides b, and if b is a unit, then p divides a. We now assume that a and b are nonzero and nonunits. Now $ab = pc$ for some $c \in D$. Let $d = pc = ab$. Since neither a nor b is a unit, it follows that d is not a unit. If c is a unit, then d is irreducible and so either a or b must be a unit, a contradiction. Therefore, c is not a unit. Since D is a UFD, there exist irreducible elements $c_1, c_2, \ldots, c_n$, $a_1, a_2, \ldots, a_m$, and $b_1, b_2, \ldots, b_r$ in D such that $c = c_1c_2 \cdots c_n$, $a = a_1a_2 \cdots a_m$, and $b = b_1b_2 \cdots b_r$. Hence, $d = pc_1c_2 \cdots c_n = a_1a_2 \cdots a_mb_1b_2 \cdots b_r$ are two expressions of d as a finite product of irreducible elements. Since D is UFD, p must be an associate of one of the irreducible elements $a_1, a_2, \ldots, a_m$, $b_1, b_2, \ldots, b_r$. If one of $a_1, a_2, \ldots, a_m$ is an associate of p, then $p \mid a$, and if one of b_1, $b_2, \ldots, b_r$ is an associate of p, then $p \mid b$. Hence, p is prime. ■

Example 13.1.11 *Consider the integral domain* $\mathbb{Z}[i\sqrt{5}] = \{a + bi\sqrt{5} \mid a, b \in \mathbb{Z}\}$. *Define*

$$N : \mathbb{Z}[i\sqrt{5}]\backslash\{0\} \to \mathbb{Z}^{\#}$$

by

$$N(a + bi\sqrt{5}) = a^2 + 5b^2.$$

We can show that $a + bi\sqrt{5}$ *is a unit if and only if* $N(a + bi\sqrt{5}) = 1$. *Let* $a + bi\sqrt{5}$, $c + di\sqrt{5}$ *be two nonzero elements of* $\mathbb{Z}[i\sqrt{5}]$. *Then* $N((a+bi\sqrt{5})(c+di\sqrt{5})) = N((ac-5bd)+i(ad+bc)\sqrt{5}) = (ac-5bd)^2+5(ad+bc)^2 = (a^2+5b^2)(c^2+5d^2) \geq (c^2+5d^2) = N((c+di\sqrt{5}))$, *where equality holds if and only if* $N((a+bi\sqrt{5})) = 1$, *i.e., if and only if* $a + bi\sqrt{5}$ *is a unit. Hence,* $\mathbb{Z}[i\sqrt{5}]$ *is a FD by Theorem* 13.1.3. *In Example* 12.3.11, *we showed that* 3 *is an irreducible element. Now* $3 \mid (2 + i\sqrt{5})(2 - i\sqrt{5})$. *Suppose* $3 \mid (2 + i\sqrt{5})$. *Then* $2 + i\sqrt{5} = 3(m + ni\sqrt{5})$ *for some* $m + ni\sqrt{5} \in \mathbb{Z}[i\sqrt{5}]$. *This implies* $2 = 3m$ *and* $1 = 3n$, *which is impossible for integers* m *and* n. *Therefore,* $3 \nmid (2 + i\sqrt{5})$. *Similarly,* $3 \nmid (2 - i\sqrt{5})$. *Thus,* 3 *is not prime in* $\mathbb{Z}[i\sqrt{5}]$. *Hence,* $\mathbb{Z}[i\sqrt{5}]$ *is not a UFD by Theorem* 13.1.10.

In this integral domain, we can also show that 2, $1 + i\sqrt{5}$, $1 - i\sqrt{5}$ *are irreducible elements and* 2 *is not an associate of any one of* $1 + i\sqrt{5}$ *and* $1 - i\sqrt{5}$. *Hence,* $6 = 2 \cdot 3 = (1 + i\sqrt{5})(1 - i\sqrt{5})$ *are two factorizations of* 6, *but there does not exist any correspondence between the irreducible factors such that the corresponding elements are associates.*

Theorem 13.1.12 *A factorization domain* D *is a UFD if and only if every irreducible element of* D *is a prime element.*

Proof. Suppose the factorization domain D is a UFD. Then by Theorem 13.1.10, every irreducible element is a prime element.

Conversely, assume that every irreducible element is a prime element in the FD D. Suppose $a = p_1p_2 \cdots p_n = q_1q_2 \cdots q_m$ are two factorizations of a as a finite product of irreducible elements. Then $p_1p_2 \cdots p_n = q_1(q_2 \cdots q_m)$ implies that $q_1 \mid (p_1p_2 \cdots p_n)$. Since q_1 is also prime, at least one of p_1, p_2, …, p_n is divisible by q_1. Let $q_1 \mid p_1$. Now p_1 and q_1 are both irreducible. Hence, $p_1 = u_1q_1$ for some unit u_1. Then $u_1q_1p_2 \cdots p_n = q_1q_2 \cdots q_m$, from which it follows by the cancelation property that $u_1p_2 \cdots p_n = q_2 \cdots q_m = q_2(q_3 \cdots q_m)$. Now $q_2 \mid (u_1p_2 \cdots p_n)$. Since q_2 is prime, q_2 does not divide u_1. Hence, q_2 divides one of $p_2, \ldots, p_n$, say, $q_2 \mid p_2$. Then $p_2 = u_2q_2$ for some unit u_2 and $u_1u_2q_2p_3 \cdots p_n = q_2 \cdots q_m$. Canceling q_2 from this relation, we obtain $u_1u_2p_3 \cdots p_n = q_3 \cdots q_m$. If $n > m$, then proceeding this way we find that $u_1u_2 \cdots u_mp_{m+1} \cdots p_n = 1$, which implies that each of p_{m+1}, …, p_n is a unit, a contradiction. If $n < m$, then we find that $u_1u_2 \cdots u_n = q_{n+1} \cdots q_m$. This implies that each of q_{n+1}, …, q_m divides a unit, which is again a contradiction. Thus, $n = m$. Also, we have shown that the corresponding irreducible factors p_i, q_i, $i = 1, 2, \ldots, n$, in the factorizations $p_1p_2 \cdots p_n$ and $q_1q_2 \cdots q_n$ are associates. Hence, D is a UFD. ■

Theorem 13.1.13 *Every PID is a UFD.*

Proof. From Lemma 13.1.6, we find that every PID satisfies ACCP. Hence, by Theorem 13.1.7, every PID is a FD. Also, by Theorem 12.3.12, every irreducible element is prime in a PID. Thus, by Theorem 13.1.12, it follows that every PID is a UFD. ■

By Theorem 12.1.9, every Euclidean domain is a PID and hence by Theorem 13.1.13, every Euclidean domain is a UFD. This result is one of the important results in factorization theory. Let us prove this result independently. First we prove the following lemma.

Lemma 13.1.14 *Let* E *be a Euclidean domain and* $a, b \in E$. *If* $a \mid b$, $b \neq 0$, *and* a *is neither a unit nor an associate of* b, *then* $v(a) < v(b)$.

Proof. Since a is not an associate of b, it follows that $b \mid a$. Hence, $a = bq + r$, where $r = 0$ or $v(r) < v(b)$. Now $b = ac$ for some $c \in E$. This implies that $r = a - bq = a - acq = a(1 - cq)$. If $1 - cq = 0$, then c is a unit and so b is an associate of a, a contradiction. Therefore, $1 - cq \neq 0$. Thus, $v(r) = v(a(1 - cq)) \geq v(a)$ and so $v(b) > v(a)$. ■

Theorem 13.1.15 *A Euclidean domain E is a unique factorization domain.*

Proof. Let v denote the Euclidean valuation of the Euclidean domain E. By induction on $v(a)$, we first show that every nonzero element a of E is either a unit or can be written as a finite product of irreducible elements. If $v(a) = v(1)$, then a is a unit. Assume that every nonzero element $b \in E$ is either a unit or expressible as a finite product of irreducible elements if $v(b) < v(a)$, where $v(a) > v(1)$ (the induction hypothesis). If a is irreducible, there is nothing to prove. Suppose that a is not irreducible. Then $a = bc$, where neither b nor c is a unit. Suppose b is an associate of a. Then $b = au$ for some unit $u \in E$. Thus, $a = bc = auc$ and so $1 = uc$, i.e., c is a unit, a contradiction. Therefore, b is not an associate of a. Similarly, c is not an associate of a. By Lemma 13.1.14, it now follows that $v(b) < v(a)$ and $v(c) < v(a)$. Thus, by our induction hypothesis, b and c are expressible as a finite product of irreducible elements of E. Hence, so is a.

The uniqueness of the factorization follows as in Theorem 13.1.12 ■

From Theorem 12.1.9, we know that every Euclidean domain is a principal ideal domain. We noted in the remark on page 187 that the converse of this result is not true. In Theorem 13.1.13, we showed that every principal ideal domain is a unique factorization domain. The converse of this result is also not true. There is a class of rings for which the converse is true. Call a complex number an **algebraic integer** if it is a root of a monic polynomial $p(x)$ in $\mathbb{Z}[x]$. The set of all algebraic integers in a finite field extension (Chapter 24) of $\mathbb{Q}$ is such a ring. However, most of these rings are not unique factorization domains. For example, the ring $\mathbb{Z}[i\sqrt{5}]$ in Example 13.1.11 is a ring in which there is no unique factorization. Here $6 = (1 - i\sqrt{5})(1 + i\sqrt{5}) = 2 \cdot 3$ are two factorizations of 6 as a product of two irreducible elements. However, the ideal $\langle 6 \rangle$ has a unique (up to order) factorization as a product of prime ideals (defined in Chapter 17), $\langle 6 \rangle = \langle 3, 1 + i\sqrt{5} \rangle \langle 3, 1 - i\sqrt{5} \rangle \langle 2, 1 + i\sqrt{5} \rangle^2$. As a matter of fact, the entire class of rings in question has the property that every ideal has a unique factorization as a product of prime ideals.

Worked-Out Exercises

♢ **Exercise 1** Show that the integral domain $\mathbb{Z}[\sqrt{10}] = \{a + b\sqrt{10} \mid a, b \in \mathbb{Z}\}$ is a FD.

Solution: Define $N : \mathbb{Z}[\sqrt{10}] \backslash \{0\} \to \mathbb{Z}^{\#}$ by for all $a + b\sqrt{10} \in \mathbb{Z}[\sqrt{10}]$,

$$N(a + b\sqrt{10}) = \left|a^2 - 10b^2\right|.$$

Now $N(a + b\sqrt{10}) = 1$ if and only if $\left|a^2 - 10b^2\right| = 1$ if and only if $(a + b\sqrt{10})(a - b\sqrt{10}) = \pm 1$ if and only if $a + b\sqrt{10}$ is a unit. Let $a + b\sqrt{10}$, $c + d\sqrt{10}$ be two nonzero elements of $\mathbb{Z}[\sqrt{10}]$. Then $N((a + b\sqrt{10})(c + d\sqrt{10})) = \left|a^2 - 10b^2\right| \left|c^2 - 10d^2\right| \geq \left|c^2 - 10d^2\right| = N((c + d\sqrt{10}))$, where equality holds if and only if $N((a + b\sqrt{10})) = 1$, i.e., if and only if $a + b\sqrt{10}$ is a unit. Hence, $\mathbb{Z}[\sqrt{5}]$ is a FD by Theorem 13.1.3.

♢ **Exercise 2** Show that in a UFD, every nonzero nonunit has only a finite number of nonassociated nontrivial factors.

Solution: Let D be a UFD. Suppose a is a nonzero nonunit element of D. Then a can be expressed uniquely as

$$a = p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k},$$

where $p_1, p_2, \ldots, p_k$ are distinct primes and $r_1, r_2, \ldots, r_k$ are positive integers. Let $d = p_1^{t_1} p_2^{t_2} \cdots p_k^{t_k}$, where $0 \leq t_i \leq r_i$, $i = 1, 2, \ldots, k$. Then d is a divisor of a. Now suppose d is any divisor of a and d is a nonunit. Then d can be expressed uniquely as $d = q_1^{t_1} q_2^{t_2} \cdots q_m^{t_m}$, where $q_1, q_2, \ldots, q_m$ are distinct primes and $t_1, t_2, \ldots, t_m$ are positive integers. Since $d \mid a$, for all $i = 1, 2, \ldots, m$, $q_i^{t_i} \mid p_j^{r_j}$ for some j, $1 \leq j \leq k$. Then $q_i \mid p_j^{r_j}$ and so $q_i \mid p_j$. Therefore, q_i is an associate of p_j. Also, we find that $t_i \leq r_j$. Thus, d is an associate of $p_1^{l_1} p_2^{l_2} \cdots p_k^{l_k}$, $0 \leq l_i \leq r_i$, $i = 1, 2, \ldots, k$. Consequently, a has only a finite number of nonassociated nontrivial divisors.

♢ **Exercise 3** Let $R = \{a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{Q}[x] \mid a_0 \in \mathbb{Z},\ n \in \mathbb{Z}^{\#}\}$. Show that R is not a UFD.

Solution: Clearly R is a subring of $\mathbb{Q}[x]$ and R contains 1. Hence, R is an integral domain. Now any unit of R is also a unit of $\mathbb{Q}[x]$. In $\mathbb{Q}[x]$, the units are the nonzero elements of $\mathbb{Q}$. Since $R \cap \mathbb{Q} = \mathbb{Z}$, it follows that 1 and -1 are the only units of R. For any nonnegative integer n, $\frac{1}{2^n} x \in R$ and $\frac{1}{2^n} x$ is not an associate of $\frac{1}{2^m} x$ when $n \neq m$. Now $x = 2^n(\frac{1}{2^n} x)$ shows that $\frac{1}{2^n} x$ is a divisor of x. Hence, x has infinite number of nontrivial

divisors in R. If R is a UFD, then x cannot have an infinite number of nontrivial divisors. Thus, R is not a UFD.

♢ **Exercise 4** In a UFD, show that the gcd of any two nonzero elements exists.

Solution: Let R be a UFD and a, b be nonzero elements of R. If one of a or b is a unit, then $\gcd(a, b) = 1$. Suppose a and b are nonunits. Then a can be expressed uniquely as

$$a = p_1^{t_1} p_2^{t_2} \cdots p_k^{t_k},$$

where $p_1, p_2, \ldots, p_k$ are irreducible elements such that p_i is not an associate of p_j when $i \neq j$ and $t_1, t_2, \ldots, t_k$ are positive integers. Similarly, b can be expressed uniquely (up to associates) as

$$b = q_1^{r_1} q_2^{r_2} \cdots q_n^{r_n},$$

where $q_1, q_2, \ldots, q_n$ are irreducible and $r_1, r_2, \ldots, r_n$ are positive integers. Now if q_1 is not an associate of any of $p_1, \ldots, p_k$, then we write $a = p_1^{t_1} \cdots p_k^{t_k} \cdot q_1^0$. Next if q_2 is not an associate of any of $p_1, p_2, \ldots, p_k$, then we write $a = p_1^{t_1} p_2^{t_2} \cdots p_k^{t_k} q_1^0 q_2^0$. But, if q_2 is an associate of one of $p_1, p_2, \ldots, p_k$, then skip q_2 and consider q_3. Continue the process for $q_3, \ldots, q_n$. We do the same thing for b. So we can write

$$\begin{aligned} a &= u_1^{n_1} u_2^{n_2} \cdots u_m^{n_m} \\ b &= u_1^{l_1} u_2^{l_2} \cdots u_m^{l_m}, \end{aligned}$$

where $u_1, u_2, \ldots, u_m$ are irreducible elements such that u_i is not an associate of u_j when $i \neq j$ and $n_1, n_2, \ldots, n_m$, $l_1, l_2, \ldots, l_m$ are nonnegative integers. Let $d = u_1^{k_1} u_2^{k_2} \cdots u_m^{k_m}$, where $k_i = \min\{n_i, l_i\}$, $i = 1, 2, \ldots, m$. Then $d \mid a$ and $d \mid b$. Let $c \mid a$ and $c \mid b$, $c \in R$. Since any irreducible divisor of c is an associate of one of $u_1, u_2, \ldots, u_m$, it follows that c must be of the form

$$c = u_1^{h_1} u_2^{h_2} \cdots u_m^{h_m},$$

where $h_i \geq 0$, and $h_i \leq n_i$, $h_i \leq l_i$, $i = 1, 2, \ldots, m$. Thus, $h_i \leq k_i$, $i = 1, 2, \ldots, m$. Hence, $c \mid d$. Thus, $d = \gcd(a, b)$.

Exercises

1. Show that $\mathbb{Z}$ satisfies the ACCP.
2. If the integral domain R satisfies the ACCP, prove that the polynomial ring $R[x]$ satisfies the ACCP.
3. Prove that an integral domain D is a UFD if and only if D satisfies the ACCP and every irreducible element is prime in D.
4. Show that the integral domains $\mathbb{Z}[i\sqrt{6}]$, $\mathbb{Z}[i\sqrt{7}]$, and $\mathbb{Z}[i\sqrt{10}]$ are factorization domains, but not unique factorization domains.
5. Let a, b be two nonzero elements of a UFD D. If $\gcd(a, b) = 1$ and $a \mid c$, $b \mid c$, prove that $ab \mid c$ in D, where $c \in D$.
6. For the following statements, write the proof if the statement is true; otherwise, give a counterexample.
   (i) Any subring of a UFD with identity is also a UFD.
   (ii) 1 and -1 are the only units of a UFD.

13.2 Factorization of Polynomials over a UFD

In this section, we show that every polynomial of degree ≥ 1 over a UFD R can be uniquely expressed as a product of irreducible polynomials over R.

Definition 13.2.1 *Let $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ be a nonzero polynomial in $R[x]$. Then the gcd$\{a_0, a_1, \ldots, a_n\}$ is called the* ***content*** *of $f(x)$.*

It is known that the gcd of $\{a_0, a_1, \ldots, a_n\}$ is not unique. If u and v are two gcd's of $\{a_0, a_1, \ldots, a_n\}$, then u and v are associates. Hence, if c_1 and c_2 are two contents of $f(x)$, then c_1 and c_2 are associates and any associate of c_1 is also a content of $f(x)$. If a and b are two elements of R such that a is an associate of b, then we write $a \sim b$.

The content of $f(x)$ is denoted by $\text{cont} f(x)$.

Definition 13.2.2 *A nonzero polynomial $f(x) \in R[x]$ is called a* ***primitive*** *polynomial if cont$f(x)$ is a unit.*

Lemma 13.2.3 *Let R be a UFD. Let $f(x)$ and $g(x)$ be two primitive polynomials in $R[x]$. Then $f(x)g(x)$ is also a primitive polynomial in $R[x]$.*

Proof. Let $f(x) = a_0+a_1x+\cdots+a_nx^n$ and $g(x) = b_0+b_1x+\cdots+b_mx^m$. Let $c_f \sim$cont$f(x)$ and $c_g \sim$cont$g(x)$. Since $f(x)$ and $g(x)$ are primitive, c_f and c_g are unit elements in R. Suppose that $f(x)g(x)$ is not a primitive polynomial. Let $f(x)g(x) = c_0+c_1x+\cdots+c_{n+m}x^{n+m}$, where $c_0 = a_0b_0$, $c_1 = a_0b_1+a_1b_0, \ldots, c_i = \sum_{j=0}^{i} a_jb_{i-j}$, where $a_j = 0$ if $j > n$, and $b_{i-j} = 0$ if $i-j > m$. Now cont$f(x)g(x)$ is not a unit. Let p be a prime element in R such that p divides cont$f(x)g(x)$. Then p divides c_i for all $i = 0, 1, \ldots, n+m$. Since c_f and c_g are unit elements, p does not divide each of $a_0, a_1, \ldots, a_n$ and also p does not divide each of $b_0, b_1, \ldots, b_m$. Let t be the smallest nonnegative integer such that p does not divide a_t. Then p divides a_i, for $i = 0, 1, \ldots, t-1$, and p does not divide a_t. Similarly, let r be the smallest nonnegative integer such that p does not divide b_r. Then p divides b_j, for $j = 0, 1, \ldots, r-1$, and p does not divide b_r. Therefore, p does not divide a_tb_r. Now $c_{t+r} = a_0b_{t+r}+ a_1b_{t+r-1}+ \cdots+ a_{t-1}b_{r+1}+ a_tb_r + a_{t+1}b_{r-1}+ \cdots + a_{t+r}b_0$, where $b_i = 0$ if $i > m$ and $a_i = 0$ if $i > n$. Now p divides a_i, for $i = 0, 1, \ldots, t-1$, p divides b_j, for $j = 0, 1, \ldots, r-1$, and p divides c_{t+r}. Hence, p divides a_tb_r, which is a contradiction. Thus, cont$f(x)g(x)$ is a unit and so $f(x)g(x)$ is a primitive polynomial. ■

Example 13.2.4 *In $\mathbb{Z}[x]$, $6x^2 + 3x - 9 = 3(2x^2 + x - 3)$. Hence, $6x^2 + 3x - 9$ is not a primitive polynomial. But $2x^2 + x - 3$ is a primitive polynomial.*

Theorem 13.2.5 *Let R be a UFD. Let $f(x)$ and $g(x)$ be two nonzero polynomials in $R[x]$. Then there exists a unit $u \in R$ such that*

$$cont(f(x)g(x)) = ucontf(x)contg(x).$$

Proof. Let c_f denote cont$f(x)$ and c_g denote cont$g(x)$. Then $f(x) = c_ff_1(x)$ and $g(x) = c_gg_1(x)$, where $f_1(x)$ and $g_1(x)$ are primitive polynomials in $R[x]$. Now cont$(f(x)g(x))$ and cont$(c_fc_gf_1(x)g_1(x))$ are associates. Since c_fc_g is a nonzero element of R, it follows that

$$\text{cont}(c_fc_gf_1(x)g_1(x))$$

and

$$c_fc_g\text{cont}(f_1(x)g_1(x))$$

are associates. By Lemma 13.2.3, cont$(f_1(x)g_1(x))$ is a unit. Hence,

$$\text{cont}(f(x)g(x)) = uc_fc_g$$

for some unit u. ■

It is known that the polynomial ring $F[x]$ over a field F is a Euclidean domain, and hence a unique factorization domain. To take advantage of this result, let us extend an integral domain R to its quotient field $Q(R)$ and establish the relationship between elements of $Q(R)[x]$ and $R[x]$.

In the remainder of the section, we let $Q(R)$ denote the quotient field of R.

Lemma 13.2.6 *Let R be a UFD. If $f(x)$ is a nonzero polynomial in $Q(R)[x]$, then there exist nonzero elements $a, b \in R$ and a primitive polynomial $f_1(x)$ in $R[x]$ such that $f(x) = ab^{-1}f_1(x)$, where b^{-1} is the inverse of b in $Q(R)[x]$.*

Proof. Let $f(x) = c_0+c_1x+\cdots+c_nx^n \in Q(R)[x]$ be a nonzero polynomial. Then $c_i \in Q(R)$, $i = 0, 1, \ldots, n$. Therefore, there exist $a_i, b_i \in R$ such that $c_i = a_ib_i^{-1}$, $b_i \neq 0$, $i = 0, 1, \ldots, n$. Now $f(x) = a_0b_0^{-1} + a_1b_1^{-1}x + \cdots + a_nb_n^{-1}x^n$. Let $b = b_0b_1\cdots b_n$. Then

$$bf(x) = a_0b_1\cdots b_n + a_1b_0b_2\cdots b_nx + \cdots + a_nb_0b_1\cdots b_{n-1}x^n \in R[x].$$

Clearly $bf(x)$ is nonzero. Let $a = \text{cont}(bf(x))$. Then $bf(x) = af_1(x)$, where cont$f_1(x)$ is a unit and $f_1(x) \in R[x]$. Hence, $f(x) = b^{-1}af_1(x)$, where $b, a \in R$ and $f_1(x)$ is a primitive polynomial in $R[x]$. ■

Lemma 13.2.7 *Let R be a UFD. Let $f(x)$ be a nonzero polynomial in $R[x]$. If $f(x) = d_1f_1(x) = d_2f_2(x)$, where $f_1(x)$ and $f_2(x)$ are primitive polynomials in $R[x]$ and $d_1, d_2 \in Q(R)$, then $d_1 = ud_2$ for some unit $u \in R$.*

Proof. Since $d_1, d_2 \in Q(R)$, we can write $d_1 = ab^{-1}$ and $d_2 = cd^{-1}$for some $a, b, c, d \in R$. Thus, $f(x) = ab^{-1}f_1(x) = cd^{-1}f_2(x)$. This implies that $adf_1(x) = cbf_2(x)$. Since $f_1(x)$ and $f_2(x)$ are primitive, $ad = ucb$ for some unit $u \in R$ by Theorem 13.2.5. Thus, $d_1 = ab^{-1} = ucd^{-1} = ud_2$. ■

Lemma 13.2.8 *Let R be a UFD. Let $f(x)$ be a nonconstant primitive polynomial in $R[x]$. Then $f(x)$ is irreducible in $R[x]$ if and only if $f(x)$ is irreducible in $Q(R)[x]$.*

Proof. Suppose $f(x)$ is irreducible in $R[x]$ and $f(x)$ is not irreducible in $Q(R)[x]$. Then there exist $h(x), g(x) \in Q(R)[x]$ such that $f(x) = h(x)g(x)$, $\deg h(x) \geq 1$, and $\deg g(x) \geq 1$. By Lemma 13.2.6, there exist $a, b, c, d \in R$ with $b \neq 0$, $d \neq 0$, and primitive polynomials $h_1(x)$, $g_1(x) \in R[x]$ such that $h(x) = ab^{-1}h_1(x)$ and $g(x) = cd^{-1}g_1(x)$. Hence, $f(x) = ab^{-1}cd^{-1}h_1(x)g_1(x)$. This implies that $bdf(x) = ach_1(x)g_1(x)$. Now $f(x)$ is primitive and so $\text{cont} f(x)$ is a unit. Thus, $\text{cont}(bdf(x)) = bdu$ for some unit u. Now

$$\begin{aligned} \text{cont}(ach_1(x)g_1(x)) &= vac\ \text{cont}(h_1(x)g_1(x)) \text{ for some unit } v \in R \\ &= v_1ac\ \text{cont}(h_1(x))\ \text{cont}(g_1(x)) \text{ for some unit } v_1 \in R \\ &= v_1acv_2v_3 \text{ for some units } v_2, v_3 \in R. \end{aligned}$$

Hence, $bd = acw$ for some unit $w \in R$. Thus, $f(x) = wh_1(x)g_1(x)$ for some unit $w \in R$. This shows that $f(x)$ is not irreducible in $R[x]$, which is a contradiction. Therefore, $f(x)$ is irreducible in $Q(R)[x]$. Conversely, let $f(x)$ be irreducible in $Q(R)[x]$. Suppose $f(x)$ is reducible in $R[x]$. Now $f(x) = rg(x)$, where $r \in R$ and r is a not a unit is impossible since $f(x)$ is primitive. Thus, there exist polynomials $f_1(x)$, $f_2(x)$ in $R[x]$ such that $\deg f_1(x) \geq 1$, $\deg f_2(x) \geq 1$, and $f(x) = f_1(x)f_2(x)$. Now $f_1(x)$ and $f_2(x)$ are also nonconstant polynomials in $Q(R)[x]$. Hence, $f(x)$ is not irreducible in $Q(R)[x]$, a contradiction. Consequently, $f(x)$ is irreducible in $R[x]$. ■

Example 13.2.9 *Consider the polynomial $4x + 4$ in $\mathbb{Q}[x]$. Now $4x + 4 = 4(x + 1)$. 4 is a unit in $\mathbb{Q}[x]$ and $x + 1$ is irreducible in $\mathbb{Q}[x]$. Hence, $4x + 4$ is irreducible in $\mathbb{Q}[x]$. But 4 is not a unit in $\mathbb{Z}[x]$. Hence, $4x + 4$ is not irreducible in $\mathbb{Z}[x]$. Also, 3 is irreducible in $\mathbb{Z}[x]$, but 3 is not irreducible in $\mathbb{Q}[x]$.*

We are now in a position to prove our main result of this section. Before proving this theorem, let us recall the following assertions concerning the polynomial ring $R[x]$ so that we can enjoy the beauty and depth of this theorem.

(i) If R is a commutative ring with 1, then $R[x]$ is a commutative ring with 1.
(ii) If R is an integral domain, then $R[x]$ is an integral domain.
(iii) If R is a field, then $R[x]$ is not a field, but $R[x]$ is a Euclidean domain.
(iv) If R is a PID, then $R[x]$ may not be a PID.

Theorem 13.2.10 *Let R be a UFD. Then $R[x]$ is a UFD.*

Proof. Let $f(x)$ be a polynomial of degree $n \geq 1$. Let $f(x) = c_f f_1(x)$, where c_f is a content of $f(x)$ and $f_1(x)$ is a primitive polynomial in $R[x]$. Now $Q(R)[x]$ is a UFD and $f_1(x) \in R[x] \subseteq Q(R)[x]$. Therefore, there exist irreducible polynomials $g_1(x), g_2(x), \ldots, g_r(x)$ in $Q(R)[x]$ such that $f_1(x) = g_1(x)g_2(x)\cdots g_r(x)$. By Lemma 13.2.7, $g_i(x) = a_ib_i^{-1}h_i(x)$, $a_i, b_i \in R$, $b_i \neq 0$, and $h_i(x)$ is a primitive polynomial in $R[x]$, $i = 1, 2, \ldots, r$. Also, by Lemma 13.2.8, $h_i(x)$ is irreducible in $R[x]$, $i = 1, 2, \ldots, r$. Hence,

$$f_1(x) = a_1a_2\cdots a_rb_1^{-1}b_2^{-1}\cdots b_r^{-1}h_1(x)\cdots h_r(x).$$

Let $a = a_1a_2\cdots a_r$ and $b = b_1b_2\cdots b_r$. Then

$$bf_1(x) = ah_1(x)\cdots h_r(x). \tag{13.1}$$

By Lemma 13.2.3, $h_1(x)\cdots h_r(x)$ is primitive. This implies that $a = ub$ for some unit $u \in R$ and so

$$f_1(x) = uh_1(x)\cdots h_r(x).$$

This shows that

$$f(x) = uc_fh_1(x)\cdots h_r(x). \tag{13.2}$$

Since an associate of an irreducible polynomial is also an irreducible polynomial, it follows that $uh_1(x)$ is irreducible. Thus, for any polynomial $f(x)$ of degree ≥ 1, there exist irreducible polynomials $g_1(x), \ldots, g_k(x)$ in $R[x]$ such that

$$f(x) = c_fg_1(x)\cdots g_k(x),$$

where $c_f = \text{cont} f(x)$. If c_f is not a unit, then there exist irreducible elements $a_1, a_2, \ldots, a_t \in R$ such that

$$f(x) = a_1a_2\cdots a_tg_1(x)\cdots g_k(x). \tag{13.3}$$

Suppose now that

$$f(x) = a_1a_2\cdots a_tg_1(x)\cdots g_k(x) = b_1b_2\cdots b_lh_1(x)\cdots h_q(x), \tag{13.4}$$

where a_i, b_j are irreducible elements in R, $i = 1, \ldots, t$, $j = 1, \ldots, l$ and

$$g_1(x), \ldots, g_k(x),\ h_1(x), \ldots, h_q(x)$$

are irreducible elements in $R[x]$. Now $a_1 a_2 \cdots a_t$ and $b_1 b_2 \cdots b_l$ are two factorizations as a product of irreducible elements in R of c_f. Therefore, by (13.4)

$$g_1(x) \cdots g_k(x) = d h_1(x) \cdots h_q(x), \tag{13.5}$$

where d is a unit in R. Now $g_1(x), \ldots, g_k(x), h_1(x), \ldots, h_q(x)$ are primitive and irreducible in $R[x]$. Hence, these polynomials are also irreducible in $Q(R)[x]$. Since $Q(R)[x]$ is a UFD, Eq. (13.5) implies that $k = q$ and there exists a one-one correspondence between $\{g_1(x), \ldots, g_k(x)\}$ and $\{h_1(x), \ldots, h_q(x)\}$ such that the corresponding factors are associates in $Q(R)[x]$ and hence by Lemma 13.2.7, they are also associates in $R[x]$. Thus, the factorization (13.4) of $f(x)$ in $R[x]$ is unique. Consequently, $R[x]$ is a UFD. ■

Corollary 13.2.11 *Let R be a UFD. The polynomial ring $R[x_1, \ldots, x_n]$ is a UFD.* ■

We see that the polynomial ring $F[x, y]$ is a unique factorization domain. However, $F[x, y]$ is not a Euclidean domain. This can be verified by showing that $F[x, y]$ is not a principal ideal ring. We ask the reader to show in the exercises that the ideal $\langle x, y\rangle$ in $F[x, y]$ is not a principal ideal.

As shown in Example 13.1.11, $\mathbb{Z}[i\sqrt{5}]$ is not a UFD. Thus, even though the polynomial ring $F[x]$ is a unique factorization domain, a ring of the form $F[c]$ need not be one. Thus, the homomorphic image of a unique factorization domain need not be a unique factorization domain.

Worked-Out Exercises

♢ **Exercise 1** Let $f(x)$ be a nonzero polynomial in $\mathbb{Z}[x]$. Show that $f(x)$ can be expressed as a product of two polynomials $g(x)$ and $h(x)$ of $\mathbb{Q}[x]$ with $\deg g(x) < \deg f(x)$ and $\deg h(x) < \deg f(x)$ if and only if there exist $g_1(x)$, $h_1(x) \in \mathbb{Z}[x]$ such that $\deg g(x) = \deg g_1(x)$, $\deg h(x) = \deg h_1(x)$, and $f(x) = g_1(x)h_1(x)$.

Solution: Suppose there exist $g(x)$ and $h(x)$ in $\mathbb{Q}[x]$ with $\deg g(x) < \deg f(x)$, $\deg h(x) < \deg f(x)$, and $f(x) = g(x)h(x)$. There exist nonzero elements $a, b, c, d \in \mathbb{Z}$ and primitive polynomials $g_2(x)$, $h_2(x) \in \mathbb{Z}[x]$ such that $g(x) = ab^{-1}g_2(x)$ and $h(x) = cd^{-1}h_2(x)$ by Lemma 13.2.6. Hence, $f(x) = ab^{-1}cd^{-1}g_2(x)h_2(x)$. This implies that $bdf(x) = acg_2(x)h_2(x)$. Let d_1 be the content of $f(x)$. Then we can write $f(x) = d_1 f_1(x)$, where $f_1(x)$ is a primitive polynomial in $\mathbb{Z}[x]$. Hence, $bdd_1 f_1(x) = acg_2(x)h_2(x)$. Now $g_2(x)h_2(x)$ is also a primitive polynomial. Then $bdd_1 = uac$ for some unit $u \in \mathbb{Z}$. This implies $bdd_1 = ac$ or $bdd_1 = -ac$. Hence, $f_1(x) = g_2(x)h_2(x)$ or $f_1(x) = -g_2(x)h_2(x)$. Let $g_1(x) = d_1 g_2(x)$. Now $f(x) = d_1 f_1(x) = d_1 g_2(x)h_2(x) = g_1(x)h_1(x)$, where $h_1(x) = h_2(x)$ or $f(x) = d_1 f_1(x) = -d_1 g_2(x)h_2(x) = g_1(x)h_1(x)$, where $h_1(x) = -h_2(x)$. Also, from the construction, it follows that $\deg g_2(x) = \deg g_1(x) = \deg g(x) < \deg f(x)$ and $\deg h_2(x) = \deg h_1(x) = \deg h(x) < \deg f(x)$. The converse is trivial.

♢ **Exercise 2** Show that $\mathbb{Z}[x]$ is a UFD, but not a PID.

Solution: Since $\mathbb{Z}$ is a UFD, $\mathbb{Z}[x]$ is a UFD by Theorem 13.2.10. (By Corollary 12.1.11, $\mathbb{Z}[x]$ is not a PID. However, here we want to show that $\mathbb{Z}[x]$ is not a PID by showing the existence of ideals in $\mathbb{Z}[x]$, which are not principal.) Consider

$$I = \langle x\rangle + \langle n\rangle,$$

where $n \in \mathbb{Z}$, $n \notin \{0, 1, -1\}$. We claim that I is not a principal ideal. Suppose $I = \langle f(x)\rangle$, where $f(x) \in \mathbb{Z}[x]$. Then $\langle n\rangle \subseteq \langle f(x)\rangle$. Therefore, $n = f(x)g(x)$ for some $g(x) \in \mathbb{Z}$. Since $\deg n = 0$, $\deg f(x) = 0$ and hence $f(x) \in \mathbb{Z}$. Let $f(x) = a \in \mathbb{Z}$. Now $\langle x\rangle \subseteq \langle a\rangle$. Then $x = ah(x)$ for some $h(x) \in \mathbb{Z}[x]$. Again by a degree argument, $\deg h(x) = 1$. Let $h(x) = a_0 + a_1 x$, where $a_0, a_1 \in \mathbb{Z}$, $a_1 \neq 0$. Then $x = a(a_0 + a_1 x)$. Hence, $1 = aa_1 \in \langle a\rangle = I = \langle x\rangle + \langle n\rangle$. Thus, $1 = xs(x) + nt(x)$ for some $s(x), t(x) \in \mathbb{Z}[x]$. Let $t(x) = t_0 + t_1 x + \cdots + t_r x^r$. Then by comparing coefficients in $1 = xs(x) + nt(x)$, we get $1 = nt_0$. Hence, n divides 1, which is a contradiction. Therefore, I is not a principal ideal.

Exercises

1. Let $f(x) \in \mathbb{Z}[x]$ be irreducible. Prove that $f(x)$ is primitive.
2. Let $f(x)$ be a nonconstant primitive polynomial in $\mathbb{Z}[x]$. Prove that if $f(x)$ is not irreducible in $\mathbb{Q}[x]$, then $f(x)$ is not irreducible in $\mathbb{Z}[x]$.
3. Show that the polynomial ring $\mathbb{Q}[x, y]$ is a UFD, but not a PID.
4. Let R be a UFD. Let $f(x)$ be a primitive polynomial in $R[x]$. Show that any nonconstant divisor of $f(x)$ is also a primitive polynomial.

13.3 Irreducibility of Polynomials

In the previous section, we proved that any polynomial of degree ≥ 1 over a UFD can be expressed as a product of irreducible polynomials. Thus, irreducible polynomials play an important role in polynomial rings. But it is not always easy to determine if a polynomial is irreducible over a UFD. In this section, we establish some criteria for irreducibility of polynomials. We first note that any polynomial of degree 1 over a field F is always irreducible. If $f(x) = ax + b \in F[x]$ with $a \neq 0$, then $x = -a^{-1}b$ is a root of $f(x)$ in F. In this connection, let us point out that a linear polynomial over a UFD D may not be irreducible in $D[x]$. For example $2x + 4 = 2(x + 2)$ is not irreducible in $\mathbb{Z}[x]$. We now consider polynomials of degree 2 and 3. For these polynomials, we can apply the following test to check irreducibility. Let F denote a field.

Theorem 13.3.1 *Let $f(x) \in F[x]$ be a polynomial of degree 2 or 3. Then $f(x)$ is irreducible over F if and only if $f(x)$ has no roots in F.*

Proof. Suppose that $\deg f(x) = 3$ and $f(x)$ is irreducible. If $f(x)$ has a root in F, say a, then $x - a$ divides $f(x)$ in $F[x]$ and so $f(x)$ is reducible over F. Conversely, suppose $f(x)$ has no roots in F. Assume that $f(x)$ is reducible. Then $f(x) = g(x)h(x)$ for some $g(x)$, $h(x) \in F[x]$, $\deg g(x) \geq 1$ and $\deg h(x) \geq 1$. Now $\deg(g(x)h(x)) = 3$. Therefore, either $\deg g(x) = 1$ and $\deg h(x) = 2$ or $\deg h(x) = 1$ and $\deg g(x) = 2$. To be specific, let $\deg g(x) = 1$ and $\deg h(x) = 2$. Then $g(x) = ax + b$ for some $a, b \in F$, $a \neq 0$. Now $-a^{-1}b \in F$ and $g(-a^{-1}b) = 0$. Thus, $-a^{-1}b$ is a root of $g(x)$ and hence $-a^{-1}b$ is a root of $f(x)$ in F. This is a contradiction to our assumption that $f(x)$ has no roots in F. Hence, $f(x)$ is irreducible over F. A similar argument can be used for the case when $\deg f(x) = 2$. ■

Example 13.3.2 *(i) Let $f(x) = x^2 + x + [1] \in \mathbb{Z}_2[x]$. Now*

$$f([0]) = [0]^2 + [0] + [1] \neq [0],$$

$$f([1]) = [1]^2 + [1] + [1] = [1] \neq [0].$$

Hence, $f(x)$ has no roots in $\mathbb{Z}_2$. Thus, by Theorem 13.3.1*, $f(x)$ is irreducible over $\mathbb{Z}_2$.*

(ii) Let $g(x) = x^3 + [2]x + [1] \in \mathbb{Z}_3[x]$. Now

$$g([0]) = [0]^3 + [2][0] + [1] \neq [0],$$

$$g([1]) = [1]^3 + [2][1] + [1] = [4] = [1] \neq [0],$$

and

$$g([2]) = [2]^3 + [2][2] + [1] = [13] = [1] \neq [0].$$

Hence, $g(x)$ has no roots in $\mathbb{Z}_3$. Thus, by Theorem 13.3.1*, $g(x)$ is irreducible over $\mathbb{Z}_3$.*

Instead of considering polynomials over an arbitrary field, let us now consider polynomials over the field $\mathbb{Q}$ of all rational numbers. By Lemma 13.2.8, a nonconstant primitive polynomial $f(x) \in \mathbb{Z}[x]$ is irreducible in $\mathbb{Q}[x]$ if and only if $f(x)$ is irreducible in $\mathbb{Z}[x]$. It is not difficult to decide whether or not a polynomial is primitive. In order to decide whether or not $f(x)$ is irreducible, we sometimes consider the corresponding polynomial in $\mathbb{Z}_p[x]$ for some prime p.

Theorem 13.3.3 *Let $f(x) = a_0 + a_1x + \cdots + a_nx^n \in \mathbb{Z}[x]$ be of degree $n > 1$. If there exists a prime p such that $\overline{f}(x) = [a_0] + [a_1]x + \cdots + [a_n]x^n$ is irreducible in $\mathbb{Z}_p[x]$ and $\deg f(x) = \deg \overline{f}(x)$, then $f(x)$ is irreducible in $\mathbb{Q}[x]$.*

Proof. Suppose $f(x)$ satisfies the given conditions of the theorem for some prime p. Suppose $f(x)$ is reducible in $\mathbb{Q}[x]$. Then there exist polynomials $g(x) = b_0 + b_1x + \cdots + b_mx^m$ and $h(x) = c_0 + c_1x + \cdots + c_kx^k$ in $\mathbb{Z}[x]$, $0 < m < n$, $0 < k < n$ such that $f(x) = g(x)h(x)$ by Worked-Out Exercise 1 (page 206). Thus, $[a_0] + [a_1]x + \cdots + [a_n]x^n = ([b_0] + [b_1]x + \cdots + [b_m]x^m)\,([c_0] + [c_1]x + \cdots + [c_k]x^k)$. Since $\deg \overline{f}(x) = \deg f(x) = n = k + m$, it follows that $[b_m][c_k] \neq 0$ in $\mathbb{Z}_p$. Hence, $[b_m] \neq [0]$ and $[c_k] \neq [0]$. Consequently, $\overline{g}(x)$ and $\overline{h}(x)$ are nonconstant polynomials in $\mathbb{Z}_p[x]$. Since the units of $\mathbb{Z}_p[x]$ are the nonzero elements of $\mathbb{Z}_p$, it follows that $\overline{g}(x)$ and $\overline{h}(x)$ are nonunits. Therefore, $\overline{f}(x)$ is not irreducible in $\mathbb{Z}_p[x]$, a contradiction. Hence, $f(x)$ is irreducible in $\mathbb{Q}[x]$. ■

Example 13.3.4 *Consider the polynomial $f(x) = \frac{5}{7}x^3 - \frac{1}{2}x + 1$ in $\mathbb{Q}[x]$. Then $14f(x) = 10x^3 - 7x + 14$. Let $f_1(x) = 10x^3 - 7x + 14$. Now in $\mathbb{Z}_3[x]$, $\overline{f_1}(x) = [10]x^3 - [7]x + [14] = x^3 - x + [2]$. Since $\overline{f_1}([0]) = [2]$, $\overline{f_1}([1]) = [2]$, $\overline{f_1}([2]) = [2]^3 - [2] + [2] = [2]$, it follows that $\overline{f_1}(x)$ has no root in $\mathbb{Z}_3[x]$. As a result $14f(x)$ is irreducible in $\mathbb{Q}[x]$. But 14 is a unit in $\mathbb{Q}[x]$. Hence, $f(x)$ is irreducible in $\mathbb{Q}[x]$.*

Let $f(x) \in \mathbb{Q}[x]$ and $\deg f(x) \geq 2$. If $f(x)$ has a root in $\mathbb{Q}$, then $f(x)$ is reducible. The following theorem will help us to see whether a polynomial $f(x) \in \mathbb{Q}[x]$ has a root in $\mathbb{Q}$.

Theorem 13.3.5 *Let $f(x) = a_0 + a_1x + \cdots + a_nx^n \in \mathbb{Z}[x]$ be of degree n and $a_0 \neq 0$. Let $\frac{u}{v} \in \mathbb{Q}$ be a root of $f(x)$, where u and v are relatively prime. Then*

$$u \mid a_0 \text{ and } v \mid a_n.$$

Proof. Since $\frac{u}{v}$ is a root of $f(x)$,

$$0 = f(\frac{u}{v}) = a_0 + a_1\frac{u}{v} + \cdots + a_n(\frac{u}{v})^n.$$

Thus,

$$0 = a_0v^n + a_1uv^{n-1} + \cdots + a_{n-1}u^{n-1}v + a_nu^n.$$

Hence,

$$v(a_0v^{n-1} + a_1uv^{n-2} + \cdots + a_{n-1}u^{n-1}) = -a_nu^n.$$

This implies that $v \mid a_nu^n$. Since u and v are relatively prime, $v \mid a_n$. Similarly, $u \mid a_0$. ■

Example 13.3.6 *Let $f(x) = 2x^3 - 7x + 1$ and $\frac{u}{v} \in \mathbb{Q}$ be a root of $f(x)$ with $\gcd(u, v) = 1$. Then $u \mid 1$ and $v \mid 2$. Hence, $u = \pm 1$ and $v = \pm 1, \pm 2$. This implies that $\frac{u}{v} = \pm 1, \pm\frac{1}{2}$. Now $f(1) \neq 0$, $f(-1) \neq 0$, $f(\frac{1}{2}) = \frac{1}{4} - \frac{7}{2} + 1 \neq 0$, and $f(-\frac{1}{2}) = -\frac{1}{4} + \frac{7}{2} + 1 \neq 0$. So we find that $f(x)$ has no root in $\mathbb{Q}$. Thus, by Theorem* 13.3.1*, $f(x)$ is irreducible in $\mathbb{Q}[x]$. Since $f(x)$is primitive, $f(x)$ is also irreducible in $\mathbb{Z}[x]$.*

Let us now give another criterion for irreducibility. This famous criterion is known as Eisenstein's irreducibility criterion.

Theorem 13.3.7 (Eisenstein's Irreducibility Criterion) *Let D be a UFD and $Q(D)$ be its quotient field. Let*

$$f(x) = a_0 + a_1x + \cdots + a_nx^n$$

be a nonconstant polynomial in $D[x]$. Suppose that D contains a prime p such that

(i) $p \mid a_i$, $i = 0, 1, \ldots, n-1$,
(ii) $p \nmid a_n$, and
(iii) $p^2 \nmid a_0$.
Then $f(x)$ is irreducible in $Q(D)[x]$.

Proof. Case 1. $f(x)$ is a primitive polynomial in $D[x]$. Under this assumption, if we can show that $f(x)$ is irreducible in $D[x]$, then by Lemma 13.2.8, it will follow that $f(x)$ is irreducible in $Q(D)[x]$. Suppose that $f(x)$ is not irreducible in $D[x]$. Then there exist polynomials

$$\begin{aligned} g(x) &= b_0 + b_1x + \cdots + b_tx^t \\ h(x) &= c_0 + c_1x + \cdots + c_kx^k \end{aligned}$$

in $D[x]$ such that $f(x) = g(x)h(x)$ and $g(x)$ and $h(x)$ are nonunits in $D[x]$. Now $n = t + k$. If $t = 0$, then $g(x) = b_0$, a nonunit element of D. Thus, $f(x) = b_0h(x)$ implies that $f(x)$ is not primitive. Therefore, $t \neq 0$. Similarly, $k \neq 0$. Hence, $0 < t < n$ and $0 < k < n$. Now from $f(x) = g(x)h(x)$, we find that $a_0 = b_0c_0$. Since p is a prime such that $p \mid a_0$ and $p^2 \nmid a_0$, it follows that p divides one of b_0, c_0, but not both. Suppose $p \mid b_0$ and $p \nmid c_0$. Since $p \nmid a_n$ and $a_n = b_tc_k$, $p \nmid b_t$ and $p \nmid c_k$. Thus, $p \mid b_0$ and $p \nmid b_t$. Let m be the smallest positive integer such that $p \nmid b_m$. Then $p \mid b_i$ for $0 \leq i < m \leq t$. Now considering the coefficient of x^m in $f(x)$ and $g(x)h(x)$, it follows that

$$a_m = b_0c_m + b_1c_{m-1} + \cdots + b_{m-1}c_1 + b_mc_0.$$

Since $p \mid b_i$, $0 \leq i < m$, we find that $p \mid (a_m - b_mc_0)$. Since $m \leq t < n$, $p \mid a_m$. Hence, $p \mid b_mc_0$ and so $p \mid b_m$ or $p \mid c_0$ since p is prime. This is a contradiction. Therefore, $f(x)$ is irreducible in $D[x]$ and hence in $Q(D)[x]$.

Case 2. $f(x)$ is not a primitive polynomial in $D[x]$. Let $d = \gcd\{a_0, a_1, \ldots, a_n\}$ in D. Then $f(x) = df_1(x)$, where $f_1(x)$ is a primitive polynomial in $D[x]$. Let $f_1(x) = d_0 + d_1x + \cdots + d_nx^n$. Then $a_i = dd_i$, for all $i = 1, 2, \ldots, n$. Since p does not divide a_n, p does not divide d. Therefore, it now follows that $p \mid d_i$, $i = 0, 1, \ldots, n-1$, $p \nmid d_n$ and $p^2 \nmid d_0$. Thus, by Case 1, $f_1(x)$ is irreducible in $Q(D)[x]$. Now d is a unit in $Q(D)$. Hence, $f(x)$ is irreducible in $Q(D)[x]$. ■

Corollary 13.3.8 *Let D be a UFD and $f(x) = a_0 + a_1x + \cdots + a_nx^n$ be a nonconstant primitive polynomial in $D[x]$. Suppose that D contains a prime p such that*
(i) $p \mid a_i$, $i = 0, 1, \ldots, n-1$,
(ii) $p \nmid a_n$, *and*
(iii) $p^2 \nmid a_0$.
Then $f(x)$ is irreducible in $D[x]$. ■

Corollary 13.3.9 *Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ be a nonconstant polynomial in $\mathbb{Z}[x]$. If there exists a prime p such that*
(i) $p \mid a_i$, $i = 0, 1, \ldots, n-1$,
(ii) $p \nmid a_n$, *and*
(iii) $p^2 \nmid a_0$,
then $f(x)$ is irreducible in $\mathbb{Q}[x]$. ■

Corollary 13.3.10 *The **cyclotomic polynomial***

$$\phi_p(x) = 1 + x + \cdots + x^{p-1} = \frac{x^p - 1}{x - 1}$$

is irreducible in $\mathbb{Z}[x]$, where p is a prime.

Proof. Since the content of $\phi_p(x)$ is 1, we find that $\phi_p(x)$ is a primitive polynomial. Suppose $\phi_p(x)$ is not irreducible in $\mathbb{Z}[x]$. Then there exist nontrivial factors $h(x)$ and $g(x)$ of $\phi_p(x)$ such that $\phi_p(x) = h(x)g(x)$. This implies that $\phi_p(x+1) = h(x+1)g(x+1)$ is a nontrivial factorization of $\phi_p(x+1)$. However,

$$\begin{aligned} \phi_p(x+1) &= \frac{(x+1)^p - 1}{(x+1) - 1} \\ &= \frac{x^p + px^{p-1} + \cdots + \binom{p}{i}x^i + \cdots + px}{x} \\ &= p + \cdots + \binom{p}{i}x^{i-1} + \cdots + px^{p-2} + x^{p-1} \end{aligned}$$

is clearly irreducible by Eisenstein's criterion. Hence, $\phi_p(x)$ is irreducible in $\mathbb{Z}[x]$.
■

Gauss is said to have placed Eisenstein at the same mathematical level as Newton and Archimedes. However, Eisenstein's influence on mathematics is considered to be small in comparison to that of the giants of mathematics.

Worked-Out Exercises

♢ **Exercise 1** Show that $f(x) = x^3 + [2]x + [4]$ is irreducible in $\mathbb{Z}_5[x]$.

Solution: $f([0]) = [4]$, $f([1]) = [7] = [2]$, $f([2]) = [3]+[4]+[4] = [1]$, $f([3]) = [2]+[1]+[4] = [2]$, $f([4]) = [4]+[3]+[4] = [1]$. Hence, $f(x)$ has no roots in $\mathbb{Z}_5$. Thus, by Theorem 13.3.1, $f(x)$ is irreducible in $\mathbb{Z}_5[x]$.

♢ **Exercise 2** Let $f(x) = x^6 + x^3 + 1 \in \mathbb{Z}[x]$. Show that $f(x)$ is irreducible over $\mathbb{Q}$.

Solution: Now $f(x+1) = x^6 + 6x^5 + 15x^4 + 21x^3 + 18x^2 + 9x + 3$. Let $p = 3$. Then by Eisenstein's criterion, $f(x+1)$ is irreducible over $\mathbb{Q}$. Hence, $f(x)$ is irreducible over $\mathbb{Q}$.

♢ **Exercise 3** Show that $f(x) = x^4 - 5x^2 + x + 1$ is irreducible in $\mathbb{Z}[x]$.

Solution: Let us first show that $f(x)$ is irreducible in $\mathbb{Q}[x]$. If $f(x)$ has a linear factor, then $f(x)$ has a root in $\mathbb{Q}$. Let $\frac{a}{b}$ (a, b are relatively prime) be a root of $f(x)$ in $\mathbb{Q}$. Then $b \mid 1$ and $a \mid 1$ by Theorem 13.3.5. Hence, $\frac{a}{b} = 1$ or -1. But $f(1) = 1 - 5 + 1 + 1 = -2 \neq 0$ and $f(-1) = 1 - 5 - 1 + 1 = -4 \neq 0$. Therefore, $f(x)$ has no linear factors in $\mathbb{Q}[x]$. Let $f(x) = (x^2 + ax + b)(x^2 + cx + d)$ in $\mathbb{Z}[x]$. Equating coefficients of powers of x, we find that

$$c + a = 0,\ d + b + ac = -5,\ ad + bc = 1,\ bd = 1.$$

Now $bd = 1$ implies that either $b = d = 1$ or $b = d = -1$. Suppose $b = d = 1$. Then $a + c = 1$. But we also have $a + c = 0$, a contradiction. Suppose $b = d = -1$. Then $ad + bc = 1$ implies that $a + c = -1$. Thus, $a + c = -1$ and $a + c = 0$, a contradiction. Hence, we find that there are no integers a, b, c, d such that $f(x) = (x^2 + ax + b)(x^2 + cx + d)$. This also implies that $f(x)$ cannot be factored as a product of two quadratic polynomials in $\mathbb{Q}[x]$ (see Worked-Out Exercise 1, page 206). Thus, $f(x)$ is irreducible in $\mathbb{Q}[x]$. Hence, by Lemma 13.2.8, $f(x)$ is irreducible in $\mathbb{Z}[x]$.

♢ **Exercise 4** Show that $f(x) = x^5 + 15x^3 + 10x + 5$ is irreducible in $\mathbb{Z}[x]$.

Solution: The content of $f(x)$ is 1. Therefore, $f(x)$ is a primitive polynomial. Now 5 is a prime integer and $5 \mid 5$, $5 \mid 10$, $5 \mid 0$, $5 \mid 15$, $5 \nmid 1$, $5^2 \nmid 5$. Hence, by Corollary 13.3.8, $f(x)$ is irreducible in $\mathbb{Z}[x]$.

♢ **Exercise 5** Give an example of a primitive polynomial which has no root in $\mathbb{Q}$, but is reducible over $\mathbb{Z}$.

Solution: Let $f(x) = x^4 + 2x^2 + 1$. This is a primitive polynomial in $\mathbb{Z}[x]$. If possible, let $\frac{a}{b}$ be a root of $f(x)$, where $a \neq 0$, $b \neq 0$ and $\gcd(a, b) = 1$. Then $a \mid 1$ and $b \mid 1$ by Theorem 13.3.5. Hence, $\frac{a}{b} = \pm 1$. But $f(1) \neq 0$ and $f(-1) \neq 0$. Therefore, $f(x)$ has no root in $\mathbb{Q}$. Since $f(x) = (x^2 + 1)(x^2 + 1)$, $f(x)$ is reducible in $\mathbb{Z}[x]$.

Exercise 6 Show that $x^2 + x + [1]$ is the only irreducible polynomial of degree 2 over $\mathbb{Z}_2$.

Solution: Any polynomial of degree 2 over $\mathbb{Z}_2$ is of the form $ax^2 + bx + c$, where $a, b, c \in \mathbb{Z}_2 = \{[0], [1]\}$. Now $a \neq [0]$. Therefore, $a = [1]$. Then x^2, $x^2 + x$, $x^2 + [1]$, and $x^2 + x + [1]$ are the only polynomials of degree 2 over $\mathbb{Z}_2$. Now $x^2 = xx$, $x^2 + x = x(x + [1])$, and $x^2 + [1] = (x + [1])(x + [1])$ showing that x^2, $x^2 + x$, and $x^2 + [1]$ are reducible. Let $f(x) = x^2 + x + [1]$. Then $f([0]) = [1] \neq 0$ and $f([1]) = [3] = [1] \neq 0$. Therefore, $f(x)$ has no root in $\mathbb{Z}_2$. Thus, $x^2 + x + [1]$ is irreducible over $\mathbb{Z}_2$.

Exercises

1. Find all irreducible polynomials of degree ≤ 2 in $\mathbb{Z}_2[x]$. Is $x^3 + [1]$ irreducible in $\mathbb{Z}_2[x]$? If not, then express it as a product of irreducible polynomials in $\mathbb{Z}_2[x]$.
2. Show that the polynomial $x^5 + x^2 + [1]$ is irreducible in $\mathbb{Z}_2[x]$. Hence, prove that $x^5 - x^2 + 9$ is irreducible in $\mathbb{Z}[x]$.
3. Show that the polynomial $x^2 + [2]x + [6]$ is reducible in $\mathbb{Z}_2[x]$ even though $x^2 + 2x + 6$ is irreducible in $\mathbb{Z}[x]$.
4. Use Eisenstein's criterion to prove that the polynomials $x^2 + 2x + 6$ and $2x^4 + 6x^3 - 9x^2 + 15$ are irreducible over $\mathbb{Z}$.
5. For $f(x) \in D[x]$, D a UFD, prove that $f(x)$ is irreducible in $D[x]$ if and only if $f(x - c)$ is irreducible in $D[x]$ for any $c \in D$.
6. Show that the polynomials $x^3 - x^2 + 1$, $x^3 - x + 1$, and $x^3 + 2x^2 + 3$ are irreducible in $\mathbb{Z}[x]$.
7. Show that the polynomial $2x^3 - x^2 + 4x - 2$ is not irreducible in $\mathbb{Z}[x]$.
8. Show that the polynomial $x^2 + \frac{1}{3}x - \frac{2}{5}$ is irreducible in $\mathbb{Q}[x]$.
9. Prove that the polynomial $f(x) = 1 - x + x^2 - x^3 + \cdots + (-1)^{p-1}x^{p-1}$ is irreducible in $\mathbb{Z}[x]$ for any prime p.
10. Let D be a UFD and $f(x) = a_0 + a_1x + \cdots + a_nx^n \in D[x]$ be of degree n and $a_0 \neq 0$. Let $uv^{-1} \in Q(D)$ be a root of $f(x)$, where $u, v \in D$ and $\gcd(u, v) = 1$. Prove that $u \mid a_0$ and $v \mid a_n$ in D.
11. Show that for any positive integer $n > 1$, $f(x) = x^n + 2$ is irreducible in $\mathbb{Z}[x]$.
12. Find all irreducible polynomials of degree 2 over the field $\mathbb{Z}_3$.
13. If $f(x)$ is an irreducible polynomial over $\mathbb{R}$, prove that either $f(x)$ is linear or $f(x)$ is quadratic.
14. Show that there are only three irreducible monic quadratic polynomials over $\mathbb{Z}_3$.
15. (i) Show that there are only 10 irreducible monic quadratic polynomials over $\mathbb{Z}_5$.
    (ii) Let p be a prime. Find the number of irreducible monic quadratic polynomials over $\mathbb{Z}_p$.

Leopold Kronecker (1823–1891) was born on December 7, 1823, in Liegnitz, Germany, to a wealthy family. He was provided with private tutoring at home. He later entered Liegnitz Gymnasium, where E. E. Kummer was his mathematics teacher. Kummer recognized his talent and encouraged him to do independent research.

In 1841, he matriculated at the University of Berlin. There he attended Dirichlet's and Steiner's mathematics lectures. He was also attracted to astronomy and in 1843 attended the University of Bonn. He returned to Berlin in 1845, the year he received his Ph.D. His thesis was on complex units.

On Kummer's nomination, Kronecker became a full member of the Berlin Academy in 1861. He was very influential at the Academy and personally helped fifteen mathematicians, including Riemann, Sylvester, Dedekind, Hermite, and Fuchs, to get various memberships.

Kronecker's primary work is in algebraic number theory. He is believed to be one of the inventors of algebraic number theory along with Kummer and Dedekind. He was the first mathematician who clearly understood Galois's work. He also proved the fundamental theorem of finite Abelian groups.

Briefly Kronecker withdrew from academic life to manage the family business. However, he continued to do mathematics as a recreation. In 1855, he returned to the academic life in Berlin. In 1880, he became editor of the *Journal für die reine and angewandte Mathematik*.

Kronecker and Weierstrass were good friends. While Weierstrass and Cantor were creating modern analysis, Kronecker's remark that "God himself made the whole numbers—everything else is the work of men" deeply affected Cantor, who was very sensitive. His remarks in opposition to Cantor's work are believed to be a factor in Cantor's nervous breakdown.

Kronecker died on December 29, 1891.

Chapter 14

Maximal, Prime, and Primary Ideals

14.1 Maximal, Prime, and Primary Ideals

In this section, we introduce certain special ideals. These ideals are motivated in large part by certain arithmetic properties of the integers. Throughout the section, we assume that the ring R contains at least two elements.

Definition 14.1.1 *An ideal P of a ring R is called **prime** if for any two ideals A and B of R, $AB \subseteq P$ implies that either $A \subseteq P$ or $B \subseteq P$.*

The following theorem gives a useful characterization of a prime ideal with the help of elements of R. Let us first recall that if A is a left ideal and B is a right ideal of a ring R, then AB is an ideal of R. Let $a \in R$. Then Ra is a left ideal of R and aR is a right ideal of R. Thus, $R(aR)$ is an ideal of R. We denote $R(aR)$ by RaR. Also, for $a \in R$, $aRa = \{ara \mid r \in R\}$.

Theorem 14.1.2 *An ideal P of a ring R is a prime ideal if and only if for all $a, b \in R$, $aRb \subseteq P$ implies that either $a \in P$ or $b \in P$.*

Proof. Suppose P is a prime ideal and $aRb \subseteq P$, where $a, b \in R$. Let $A = RaR$ and $B = RbR$. Then A and B are ideals of R. Also, $AB = (RaR)(RbR) \subseteq R(aRb)R \subseteq RPR \subseteq P$. Since P is a prime ideal, it follows that either $A \subseteq P$ or $B \subseteq P$. Suppose $A \subseteq P$. Now $\langle a \rangle^3 \subseteq RaR = A \subseteq P$. Since P is a prime ideal, $\langle a \rangle \subseteq P$ and so $a \in P$. Similarly, if $B \subseteq P$, then $b \in P$. Thus, either $a \in P$ or $b \in P$. Conversely, suppose that the ideal P satisfies the given condition of the theorem. Let A and B be two ideals of R such that $AB \subseteq P$. Suppose that $A \not\subseteq P$. Then there exists $a \in A$ such that $a \notin P$. Let $b \in B$. Now $aRb = (aR)b \subseteq AB \subseteq P$. This implies that $a \in P$ or $b \in P$. But $a \notin P$. Therefore, $b \in P$. Hence, $B \subseteq P$. ■

Corollary 14.1.3 *Let R be a commutative ring. An ideal P of R is a prime ideal if and only if for all $a, b \in R$, $ab \in P$ implies that either $a \in P$ or $b \in P$.* ■

Example 14.1.4 *In the ring $\mathbb{Z}$ of integers, the ideal $P = \{3k \mid k \in \mathbb{Z}\}$ is a prime ideal. For, $ab \in P$ if and only if ab is divisible by 3 if and only if a is divisible by 3 or b is divisible by 3 (since 3 is prime) if and only if a is a multiple of 3 or b is a multiple of 3 if and only if $a \in P$ or $b \in P$. In $\mathbb{Z}$, the ideal $J = \{6k \mid k \in \mathbb{Z}\}$ is not a prime ideal since $3 \cdot 2 = 6 \in J$, but $3 \notin J$ and $2 \notin J$.*

Theorem 14.1.5 *Let R be a PID and P be a nonzero ideal of R. Then P is prime and $P \neq R$ if and only if P is generated by a prime element.*

Proof. Let R be a PID and $P = \langle p \rangle$ be a nonzero proper prime ideal of R. Then $p \neq 0$. Since $P \neq R$, p is not a unit. Let $a, b \in R$ be such that $p \mid ab$. Then $ab = pc$ for some $c \in R$. Hence, $ab \in P$. Since P is a prime ideal, either $a \in P$ or $b \in P$. Therefore, either $p \mid a$ or $p \mid b$. Thus, p is a prime element. Conversely, suppose that $P = \langle p \rangle$ is a nonzero ideal of R such that p is a prime element. Since p is not a unit, $P \neq R$. Let a, b be two elements of R such that $ab \in P$. Then $p \mid ab$. Since p is a prime element, either $p \mid a$ or $p \mid b$. Therefore, either $a \in P$ or $b \in P$. Hence, P is a prime ideal of R. ■

As a consequence of Theorem 14.1.5 and Theorem 12.1.9, the prime ideals of $\mathbb{Z}$ are precisely those ideals generated by primes and the ideals $\{0\}$ and $\mathbb{Z}$. Also, by Theorem 12.3.16, the prime ideals in the polynomial ring $F[x]$ over a field F are those ideals generated by irreducible polynomials and the ideals $\{0\}$ and $F[x]$.

Definition 14.1.6 *Let R be a ring and M be a (left, right) ideal of R. Then M is called a* ***maximal (left, right) ideal*** *of R if $M \neq R$ and there does not exist any (left, right) ideal I of R such that $M \subset I \subset R$.*

Theorem 14.1.7 *Let R be a commutative ring with 1. Then every maximal ideal of R is a prime ideal of R.*

Proof. Let I be a maximal ideal of R and a and b be two elements of R such that $ab \in I$ and $a \notin I$. Now $\langle I, a\rangle = \{u + ra \mid u \in I,\ r \in R\}$ is the ideal generated by $I \cup \{a\}$. Since $a \notin I$, $I \subset \langle I, a\rangle$. Also, since I is a maximal ideal, $\langle I, a\rangle = R$. Thus, there exist $u \in I$ and $r \in R$ such that $1 = u + ra$. This implies that $b = ub + rab \in I$. Hence, I is a prime ideal. ■

The converse of the above theorem is not true, as shown by the following examples.

Example 14.1.8 *In the ring $\mathbb{Z}$ of integers, $\{0\}$ is a prime ideal, but not a maximal ideal.*

Example 14.1.9 *Let $R = \{(a, b) \mid a, b \in \mathbb{Z}\}$. Then $(R, +, \cdot)$ is a ring, where $+$ and $\cdot$ are defined by*

$$\begin{aligned}(a, b) + (c, d) &= (a + c, b + d),\\ (a, b) \cdot (c, d) &= (ac, bd)\end{aligned}$$

for all $a, b, c, d \in \mathbb{Z}$. Let $I = \{(a, 0) \mid a \in \mathbb{Z}\}$. Then I is a prime ideal of R, but not a maximal ideal since $I \subset \langle I, (0, 2)\rangle \subset R$.

Theorem 14.1.10 *Let R be a principal ideal domain. Then a nonzero ideal P $(\neq R)$ of R is prime if and only if it is maximal.*

Proof. Suppose P $(\neq R)$ is a nonzero prime ideal. By Theorem 14.1.5, $P = \langle p\rangle$ for some prime element $p \in R$. We now show that there is no ideal I of R such that $P \subset I \subset R$. Suppose I is an ideal of R such that $P \subset I$. Since $P \neq I$, there exists an element $a \in I$ such that $a \notin P$. Then a and p are relatively prime and so there exist s, $t \in R$ such that $1 = sa + tp$. Since $sa \in I$ and $tp \in P \subset I$, we must have $1 \in I$. This implies that $I = R$. Hence, P is maximal. ■

We now give characterizations of prime ideals and maximal ideals in a commutative ring with identity by the quotient rings of the ideals.

Theorem 14.1.11 *Let R be a commutative ring with 1 and P be an ideal of R such that $P \neq R$. Then P is a prime ideal if and only if R/P is an integral domain.*

Proof. Let P be a prime ideal of R. Since R is a commutative ring with 1, the quotient ring R/P is also a commutative ring with 1. Now $P \neq R$ and so the identity element $1 + P$ of R/P is different from the zero element $0 + P$. Let us now show that R/P has no zero divisors. Let $a + P,\ b + P \in R/P$, and $(a + P)(b + P) = 0 + P$. Then $ab + P = 0 + P$, which implies that $ab \in P$. Since P is a prime ideal, either $a \in P$ or $b \in P$, i.e., either $a + P = 0 + P$ or $b + P = 0 + P$. Thus, R/P has no zero divisors. This implies that R/P is an integral domain. Conversely, suppose R/P is an integral domain. Let $ab \in P$. Then $0 + P = ab + P = (a + P)(b + P)$, whence $a + P = 0 + P$ or $b + P = 0 + P$. Thus, $a \in P$ or $b \in P$ and so P is a prime ideal. ■

Theorem 14.1.12 *Let R be a commutative ring with 1 and M be an ideal of R. Then M is a maximal ideal if and only if R/M is a field.*

Proof. Suppose that M is a maximal ideal. Since R is a commutative ring with 1, R/M is a commutative ring with 1. For all $a \in R$, let $\overline{a}$ denote the coset $a + M$ in R/M. Let $\overline{a} \in R/M$ be such that $\overline{a} \neq \overline{0}$. Then $a \notin M$. Hence, the ideal $\langle M, a\rangle$ generated by $M \cup \{a\}$ properly contains M. Since M is a maximal ideal, we have $\langle M, a\rangle = R$. This implies that there exist $m \in M$ and $r \in R$ such that $m + ra = 1$. Thus, $\overline{m} + \overline{ra} = \overline{1}$ and so $\overline{ra} = \overline{1}$. Hence, $\overline{a}$ has an inverse. This shows that every nonzero element of R/M is a unit and so R/M is a field. Conversely, suppose R/M is a field. Since R/M is a field, $R \neq M$. Let I be an ideal of R such that $M \subset I \subseteq R$. There exists $a \in I$ such that $a \notin M$. Then $\overline{a} \neq \overline{0}$ and so there exists $\overline{r} \in R/M$ such that $\overline{ar} = \overline{1}$. Thus, $(a + M)(r + M) = 1 + M$, which implies $1 - ar \in M$. Hence, $1 = m + ar$ for some $m \in M$. Thus, $1 = m + ar \in M + I \subseteq I$. This implies that $I = R$. Therefore, M is maximal. ■

As a consequence of Theorems 12.1.9 and 14.1.10, the maximal ideals of $\mathbb{Z}$ are precisely those ideals generated by primes. Also, by Theorem 12.3.16, the maximal ideals in the polynomial ring $F[x]$ over a field F are those ideals generated by irreducible polynomials.

Example 14.1.13 *Consider the polynomial ring $R[x, y]$ over an integral domain R. Then $R[x, y]/\langle x\rangle \simeq R[y]$ and $R[x, y]/\langle y\rangle \simeq R[x]$, which are integral domains. Thus, $\langle x\rangle$ and $\langle y\rangle$ are prime ideals. Since $R[x, y]/\langle x\rangle$ and $R[x, y]/\langle y\rangle$ are not fields, $\langle x\rangle$ and $\langle y\rangle$ are not maximal ideals.*

Example 14.1.14 *Consider $\mathbb{E}$, the ring of even integers. The ideal $\langle 4\rangle$ is maximal, but not prime in $\mathbb{E}$ since $2 \cdot 2 \in \langle 4\rangle$, but $2 \notin \langle 4\rangle$. Note that $\mathbb{E}$ is commutative without identity.*

We now show the existence of maximal ideals in certain rings. In order to accomplish this, we require Zorn's lemma.

Theorem 14.1.15 *Let R be a commutative ring with 1. Then every proper ideal of R is contained in a maximal ideal of R.*

Proof. Let I be a proper ideal of R and set $\mathcal{A} = \{J \mid I \subseteq J,\ J \text{ is a proper ideal of } R\}$. Since $I \in \mathcal{A}$, $\mathcal{A} \neq \emptyset$. Also, $\mathcal{A}$ is a partially ordered set, where the partial order $\leq$ is the usual set inclusion. We now show that any chain in $\mathcal{A}$ has an upper bound in $\mathcal{A}$. Let $\mathcal{C} = \{J_\alpha \mid \alpha \in K\}$ be a chain in $\mathcal{A}$. Since $I \subseteq J_\alpha$ for all α, $I \subseteq \cup_\alpha J_\alpha$. Let $a, b \in \cup_\alpha J_\alpha$. Then $a \in J_\alpha$ and $b \in J_\beta$ for some α, β. Since $\mathcal{C}$ is a chain, either $J_\alpha \subseteq J_\beta$ or $J_\beta \subseteq J_\alpha$, say, $J_\alpha \subseteq J_\beta$. Thus, $a, b \in J_\beta$. Since J_β is an ideal of R, $a - b \in J_\beta \subseteq \cup_\alpha J_\alpha$. Let $r \in R$. Then $ra \in J_\alpha \subseteq \cup_\alpha J_\alpha$, whence $\cup_\alpha J_\alpha$ is an ideal of R. Now $\cup_\alpha J_\alpha \neq R$ else $1 \in J_\alpha$ for some α, which is impossible since $J_\alpha \neq R$. Hence, $\cup_\alpha J_\alpha \in \mathcal{A}$, which is clearly an upper bound of $\mathcal{C}$ and so by Zorn's lemma, $\mathcal{A}$ has a maximal element, say, M. We now show that M is a maximal ideal. If there exists an ideal J of R such that $M \subset J \subset R$, then $J \in \mathcal{A}$ and so M is not maximal in $\mathcal{A}$, a contradiction. Thus, no such J exists and so M is a maximal ideal. ■

Corollary 14.1.16 *Let R be a commutative ring with 1 and $a \in R$. Then a is in a maximal ideal of R if and only if a is not a unit.*

Proof. Suppose a is not a unit. Then $\langle a\rangle \subset R$ else $1 = ra$ for some r. By Theorem 14.1.15, there exists a maximal ideal M such that $\langle a\rangle \subseteq M$. Now $a \in \langle a\rangle \subseteq M$. Conversely, suppose $a \in M$, where M is a maximal ideal. If a is a unit, then $1 = a^{-1}a \in M$ and so $M = R$, a contradiction. ■

Corollary 14.1.17 *Let R be a commutative ring with 1. Then R has a maximal ideal.*

Proof. In R, $\{0\}$ is a proper ideal. Hence, by Theorem 14.1.15, there exists a maximal ideal M of R such that $\{0\} \subseteq M$. ■

The fundamental theorem of arithmetic says that any integer n has a prime factorization $n = p_1^{e_1} \cdots p_s^{e_s}$, where $p_1, \ldots, p_s$ are primes and $e_1, \ldots, e_s$ are positive integers. The ideals $\langle p_i\rangle$ are prime ideals of $\mathbb{Z}$. The ideals $\langle p_i^{e_i}\rangle$ are also special ideals of $\mathbb{Z}$. Their study is motivated in part by the fundamental theorem of arithmetic.

Definition 14.1.18 *Let R be a commutative ring and Q be an ideal of R. Then Q is called a **primary ideal** if for all $a, b \in R$, $ab \in Q$ and $a \notin Q$ implies that there exists a positive integer n such that $b^n \in Q$.*

From the definition of primary ideal, it follows immediately that every prime ideal in a commutative ring is a primary ideal. Now in the ring $\mathbb{Z}$, for any prime integer p, the ideal $\langle p^n\rangle$ contains p^n but not p, where n is a positive integer and $n \geq 2$. Hence, $\langle p^n\rangle$ is not a prime ideal. The following example shows that $\langle p^n\rangle$ is a primary ideal.

Example 14.1.19 *Let p be a prime in $\mathbb{Z}$ and n be a positive integer. We show that $\langle p^n\rangle$ is a primary ideal. Let $ab \in \langle p^n\rangle$ and $a \notin \langle p^n\rangle$. Then there exists $r \in \mathbb{Z}$ such that $ab = rp^n$. Since p^n does not divide a, $p \mid b$ and so $b = qp$ for some $q \in \mathbb{Z}$. Thus, $b^n = q^n p^n$ and so $b^n \in \langle p^n\rangle$.*

Example 14.1.20 *Let $p(x)$ be irreducible in $F[x]$, F a field, and n be a positive integer. Then $\langle p(x)^n\rangle$ is a primary ideal by an argument entirely similar to the one used in Example* 14.1.19.

Definition 14.1.21 *Let R be a commutative ring and I be an ideal of R. Then the **radical** of I, denoted by $\sqrt{I}$, is defined to be the set*

$$\sqrt{I} = \{a \in R \mid a^n \in I \text{ for some positive integer } n\}.$$

Theorem 14.1.22 *Let Q be an ideal of a commutative ring R. Then*
(i) $\sqrt{Q}$ is an ideal of R and $\sqrt{Q} \supseteq Q$,
(ii) if Q is a primary ideal, then $\sqrt{Q}$ is a prime ideal.

Proof. (i) Clearly $\sqrt{Q} \supseteq Q$. Let $a, b \in \sqrt{Q}$. Then there exist positive integers n, m such that $a^n, b^m \in Q$. Thus, $(a-b)^{n+m} \in Q$ and so $a - b \in \sqrt{Q}$. Let $r \in R$. Then $(ra)^n = r^n a^n \in Q$ and so $ra \in \sqrt{Q}$. Hence, $\sqrt{Q}$ is an ideal of R.

(ii) Let a, $b \in R$ be such that $ab \in \sqrt{Q}$ and $a \notin \sqrt{Q}$. There exists a positive integer n such that $a^n b^n = (ab)^n \in Q$. But $a^n \notin Q$. Since Q is primary, there exists a positive integer m such that $b^{nm} = (b^n)^m \in Q$. Therefore, $b \in \sqrt{Q}$ and so $\sqrt{Q}$ is prime. ■

Definition 14.1.23 *Let Q be a primary ideal of a commutative ring R. Then the radical $P = \sqrt{Q}$ of Q is called the **associated prime ideal** of Q and Q is called a **primary ideal belonging to** (or **primary for**) the prime ideal P.*

Example 14.1.24 *Let i be a positive integer. In $\mathbb{Z}$, we show that $\langle p^i \rangle$ is primary for $\langle p \rangle$, where p is a prime. It suffices to show that $\langle p \rangle = \sqrt{\langle p^i \rangle}$. Let $a \in \sqrt{\langle p^i \rangle}$. Then there exists a positive integer n such that $a^n \in \langle p^i \rangle$. Therefore, $a^n = rp^i$ for some $r \in \mathbb{Z}$. This implies that $p \mid a$ and so $a \in \langle p \rangle$. Hence, $\sqrt{\langle p^i \rangle} \subseteq \langle p \rangle$. Let $a \in \langle p \rangle$. Then there exists $t \in \mathbb{Z}$ such that $a = tp$. This implies that $a^i = t^i p^i \in \langle p^i \rangle$ and so $a \in \sqrt{\langle p^i \rangle}$. Thus, $\langle p \rangle \subseteq \sqrt{\langle p^i \rangle}$.*

In $F[x]$ (F a field), a similar argument shows that $\langle p(x)^i \rangle$ is primary for $\langle p(x) \rangle$, where $p(x)$ is irreducible and $\langle p(x) \rangle = \sqrt{\langle p(x)^i \rangle}$.

Theorem 14.1.25 *Let Q and P be ideals of a commutative ring R. Then Q is primary and $P = \sqrt{Q}$ if and only if*

(i) $Q \subseteq P \subseteq \sqrt{Q}$ and

(ii) $ab \in Q$, $a \notin Q$ implies $b \in P$.

Proof. The necessity of (i) and (ii) is immediate. Suppose (i) and (ii) hold. Let $ab \in Q$, $a \notin Q$. Then $b \in P \subseteq \sqrt{Q}$ and so there exists a positive integer n such that $b^n \in Q$, whence Q is primary. We now show that $P = \sqrt{Q}$. Let $b \in \sqrt{Q}$. Then there exists a positive integer n such that $b^n \in Q \subseteq P$. Let n be the smallest positive integer such that $b^n \in Q$. If $n = 1$, then $b \in P$. So assume that $n \geq 2$. Then $bb^{n-1} \in Q$ and $b^{n-1} \notin Q$ implies that $b \in P$. Hence, $\sqrt{Q} \subseteq P$ and so $P = \sqrt{Q}$. ■

We now show that every primary ideal I of a commutative ring R can be characterized with the help of some properties of the quotient ring R/I.

Theorem 14.1.26 *Let R be a commutative ring and I be an ideal of R. Then I is a primary ideal if and only if every zero divisor of R/I is nilpotent.*

Proof. First suppose that I is a primary ideal. Let $a + I$ be a zero divisor in R/I. Then there exists an element $b + I \in R/I$, $b + I \neq I$, such that $(a + I)(b + I) = I$. Now $ab \in I$ and $b \notin I$. Since I is a primary ideal, it follows that $a^n \in I$ for some positive integer n. Hence, $(a + I)^n = a^n + I = I$, showing that $a + I$ is nilpotent.

Conversely, suppose that every zero divisor of R/I is nilpotent. Let $a, b \in R$ be such that $ab \in I$ and $a \notin I$. Then $a+I \neq I$. Now $(a+I)(b+I) = ab+I = I$. If $b+I = I$, then $b \in I$. Suppose $b+I \neq I$. This implies that $b+I$ is a zero divisor and so is nilpotent. Therefore, there exists a positive integer n such that $b^n + I = (b + I)^n = I$. Thus, $b^n \in I$. Consequently, I is a primary ideal. ■

Consider $\mathbb{Z}$. For the prime factorization of an integer n, $n = p_1^{e_1} \cdots p_s^{e_s}$, we have

$$\langle n \rangle = \langle p_1^{e_1} \rangle \cdots \langle p_s^{e_s} \rangle = \langle p_1^{e_1} \rangle \cap \cdots \cap \langle p_s^{e_s} \rangle$$

and $\sqrt{\langle p_i^{e_i} \rangle} = \langle p_i \rangle$, $i = 1, 2, \ldots, s$. However, in the polynomial ring $\mathbb{Z}[x, y]$, it can be shown that the ideal $\langle x^2, xy, 2 \rangle$ is an intersection of primary ideals, but not a product of primary ideals. These concepts involving prime and primary ideals are used in the study of nonlinear equations. For example, consider the following nonlinear equations:

$$\begin{aligned} x^2 - y &= 0 \\ x^2 z &= 0. \end{aligned}$$

In the polynomial ring $\mathbb{R}[x, y]$, let $I = \langle x^2 - y, x^2 z \rangle$. It can be shown that $\langle x^2 - y, z \rangle$ and $\langle x^2, y \rangle$ are primary ideals and that $I = \langle x^2, y \rangle \cap \langle x^2 - y, z \rangle$. In fact, it can be shown in any polynomial ring $F[x_1, \ldots, x_n]$ over a field F that every ideal is a finite intersection of primary ideals. This latter result is a type of fundamental theorem of arithmetic for ideals. It can also be shown that $\sqrt{\langle x^2 - y, z \rangle} = \langle x^2 - y, z \rangle$ and $\sqrt{\langle x^2, y \rangle} = \langle x, y \rangle$. The solution to the above system of equations is

$$\{(x, x^2, 0) \mid x \in \mathbb{R}\} \cup \{(0, 0, z) \mid z \in \mathbb{R}\}.$$

The ideal $\langle x^2 - y, z \rangle$ corresponds to $\{(x, x^2, 0) \mid x \in R\}$, while the ideal $\langle x, y \rangle$ corresponds to $\{(0, 0, z) \mid z \in \mathbb{R}\}$.

We conclude this section by mentioning the following differences between the ideals of $\mathbb{Z}$ and $\mathbb{Z}[x]$.

1. In the ring $\mathbb{Z}$, every ideal is a principal ideal, but in $\mathbb{Z}[x]$ there exist ideals (for example, $\langle x, 2\rangle$), which are not principal.
2. In the ring $\mathbb{Z}$, a nontrivial ideal is a prime ideal if and only if it is a maximal ideal. In the ring $\mathbb{Z}[x]$, there are prime ideals (for example $\langle x\rangle$), which are not maximal.
3. In the ring $\mathbb{Z}$, a nontrivial ideal I is a primary ideal if and only if $I = \langle p^n\rangle$ for some prime p and for some positive integer n. Hence, in $\mathbb{Z}$, if I is a primary ideal, then I is expressible as some power of its associated prime ideal. In $\mathbb{Z}[x]$, this is not true, as $\langle x, 4\rangle$ is a primary ideal with $\langle x, 2\rangle$ as its associated prime ideal, but $\langle x, 4\rangle \neq \langle x, 2\rangle^n$ for any $n \geq 1$.

Worked-Out Exercises

♢ **Exercise 1** Let R be an integral domain. Prove that if every ideal of R is a prime ideal, then R is a field.

Solution: Let $0 \neq a \in R$. Then a^2R is an ideal of R and hence it is a prime ideal. Now $a^2 \in a^2R$. Since a^2R is a prime ideal, $a \in a^2R$. Thus, $a = a^2b$ for some $b \in R$. Then $a(1 - ab) = 0$. Since R is an integral domain and $a \neq 0$, $1 - ab = 0$ and so $ab = 1$, proving that a is a unit. Hence, R is a field.

♢ **Exercise 2** Let R be a commutative ring with 1. Suppose that $\langle x\rangle$ is a prime ideal of $R[x]$. Show that R is an integral domain.

Solution: Since $\langle x\rangle$ is a prime ideal $R[x]/\langle x\rangle$ is an integral domain. Since $R[x]/\langle x\rangle \simeq R$, R is an integral domain.

Exercise 3 Let R be a commutative ring and I be an ideal of R. Let P be a prime ideal of I. Show that P is an ideal of R.

Solution: Let $a \in P \subseteq I$ and $r \in R$. Then $rar \in I$. Therefore, $a(rar) \in P$ and so $(ar)^2 \in P$. Since P is a prime ideal of I, $ar \in P$. Hence, P is an ideal of R.

Exercise 4 Show that a proper ideal I of a ring R is a maximal ideal if and only if for any ideal A of R either $A \subseteq I$ or $A + I = R$.

Solution: Suppose I is a maximal ideal of R and let A be any ideal of R. If $A \not\subseteq I$, then $A + I$ is an ideal of R such that $I \subset A + I$. Since I is maximal, it follows that $A + I = R$.

Conversely, assume that the proper ideal I satisfies the given condition. Let J be an ideal of R such that $I \subset J$. Now $J \not\subseteq I$. Therefore, $I + J = R$. But $I + J = J$. Thus, $J = R$. Hence, I is a maximal ideal of R.

♢ **Exercise 5** Let R be a PID which is not a field. Prove that any nontrivial ideal I of R is a maximal ideal if and only if it is generated by an irreducible element.

Solution: Since R is not a field, there exists an element $0 \neq a \in R$ such that a is not a unit. Then $\langle 0\rangle \subset \langle a\rangle \subset R$. Therefore, $\langle 0\rangle$ is not a maximal ideal. Let I be a maximal ideal of R. Then $I \neq \{0\}$ and $I = \langle p\rangle$ for some $p \in R$, where p is irreducible by Theorem 14.1.5 and Corollary 12.3.13. Conversely, let $I = \langle p\rangle$ and p be irreducible. Let $I \subset J \subseteq R$. Since R is a PID, $J = \langle a\rangle$ for some $a \in R$. Since $p \in \langle a\rangle$, a divides p. Thus, $p = ab$ for some $b \in R$. Since p is irreducible, either a is a unit or b is a unit. If b is a unit, then $a = pb^{-1} \in \langle p\rangle$. Thus, $J \subseteq I$, which is a contradiction. Hence, a is a unit and so $J = R$. Thus, I is a maximal ideal.

♢ **Exercise 6** Show that the ideal $\langle x\rangle$ in $\mathbb{Z}[x]$ is a prime ideal, but not a maximal ideal.

Solution: Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ be two elements in $\mathbb{Z}[x]$ such that $f(x)g(x) \in \langle x\rangle$. Then $a_0b_0 = 0$. Thus, either $a_0 = 0$ or $b_0 = 0$. Hence, either $f(x) \in \langle x\rangle$ or $g(x) \in \langle x\rangle$, showing that $\langle x\rangle$ is a prime ideal. Now the ideal $\langle x, 2\rangle$ of $\mathbb{Z}[x]$ is such that $\langle x\rangle \subset \langle x, 2\rangle \subset \mathbb{Z}[x]$. Hence, $\langle x\rangle$ is not a maximal ideal.

♢ **Exercise 7** Let R be a commutative ring with 1. Let A and B be two distinct maximal ideals of R. Show that $AB = A \cap B$.

Solution: Since $AB \subseteq A$ and $AB \subseteq B$, $AB \subseteq A \cap B$. Since A and B are distinct maximal ideals, there exists $b \in B$ such that $b \notin A$. Then $\langle A, b\rangle = \{a + br \mid a \in A, r \in R\}$ is an ideal of R such that $A \subset \langle A, b\rangle$. Since A is maximal, $\langle A, b\rangle = R$. This implies that $1 = a + br$ for some $a \in A$ and $r \in R$. Let $x \in A \cap B$. Then $x = x1 = xa + xbr = xa + (xb)r \in AB$. Hence, $A \cap B \subseteq AB$. Thus, $AB = A \cap B$.

♢ **Exercise 8** Let $f(x) = x^5 + 12x^4 + 9x^2 + 6$. Show that the ideal $I = \langle f(x)\rangle$ is maximal in $\mathbb{Z}[x]$.

Solution: I will be a maximal ideal if we can prove that $f(x)$ is an irreducible polynomial in $\mathbb{Z}[x]$. The content of $f(x)$ is 1. Hence, $f(x)$ is a primitive polynomial in $\mathbb{Z}[x]$. Also, for the prime 3, we find that $3 \mid 6$, $3 \mid 9$, $3 \mid 12$, $3 \nmid 1$, $3^2 \nmid 6$. Hence, $f(x)$ is irreducible in $\mathbb{Z}[x]$, by Eisenstein's criterion.

♢ **Exercise 9** (a) Find all maximal ideals of the ring $\mathbb{Z}_6$.

(b) ind all ideals and all maximal ideals of the ring $\mathbb{Z}_8$.

Solution: (a) The mapping $\beta : \mathbb{Z} \to \mathbb{Z}_6$ defined by $\beta(n) = [n]$ is a homomorphism of $\mathbb{Z}$ onto $\mathbb{Z}_6$ and Ker $\beta = 6\mathbb{Z}$. If I is any ideal of $\mathbb{Z}_6$, then there exists a unique ideal A of $\mathbb{Z}$ such that Ker $\beta \subseteq A$ and $\beta(A) = I$. Now $\mathbb{Z}$, $2\mathbb{Z}$, $3\mathbb{Z}$, and $6\mathbb{Z}$ are the only ideals of $\mathbb{Z}$ which contain $6\mathbb{Z}$. Also, $\beta(\mathbb{Z}) = \mathbb{Z}_6$, $\beta(2\mathbb{Z}) = \{[0], [2], [4]\}$, $\beta(3\mathbb{Z}) = \{[0], [3]\}$, and $\beta(6\mathbb{Z}) = \{[0]\}$. Hence, $\{[0], [2], [4]\}$ and $\{[0], [3]\}$ are the only maximal ideals of $\mathbb{Z}_6$ since $2\mathbb{Z}$ and $3\mathbb{Z}$ are maximal ideals of $\mathbb{Z}$.

(b) The mapping $\beta : \mathbb{Z} \to \mathbb{Z}_8$ defined by $\beta(n) = [n]$ is an epimorphism of rings and Ker $\beta = 8\mathbb{Z}$. Now $\mathbb{Z}$, $2\mathbb{Z}$, $4\mathbb{Z}$, and $8\mathbb{Z}$ are the only ideals of $\mathbb{Z}$ which contain $8\mathbb{Z}$. Also, $\beta(\mathbb{Z}) = \mathbb{Z}_8$, $\beta(2\mathbb{Z}) = \{[0], [2], [4], [6]\}$, $\beta(4\mathbb{Z}) = \{[0], [4]\}$, and $\beta(8\mathbb{Z}) = \{[0]\}$. Hence, the ideals of $\mathbb{Z}_8$ are $\mathbb{Z}_8$, $\{[0], [2], [4], [6]\}$, $\{[0], [4]\}$, and $\{[0]\}$. Now $\{[0]\} \subset \{[0], [4]\} \subset \{[0], [2], [4], [6]\} \subset \mathbb{Z}_8$. This implies that $\mathbb{Z}_8$ has only one maximal ideal, which is $\{[0], [2], [4], [6]\}$.

♢ **Exercise 10** Show that $\langle x^2 \rangle$ is a primary ideal in $\mathbb{Z}[x]$ with $\langle x \rangle$ as its associated prime ideal.

Solution: Let $f(x) = a_0 + a_1x + \cdots + a_nx^n$ and $g(x) = b_0 + b_1x + \cdots + b_mx^m$ be two elements in $\mathbb{Z}[x]$ such that $f(x)g(x) \in \langle x^2 \rangle$ and $f(x) \notin \langle x^2 \rangle$. Then $f(x)g(x) = x^2h(x)$ for some $h(x) \in \mathbb{Z}[x]$. Hence, $a_0b_0 = 0$ and $a_0b_1 + a_1b_0 = 0$. Since $f(x) \notin \langle x^2 \rangle$, it follows that either $a_0 \neq 0$ or $a_1 \neq 0$. If $a_0 \neq 0$, then $b_0 = 0$ and $b_1 = 0$ and so $g(x) \in \langle x^2 \rangle$. If $a_0 = 0$, then $a_1 \neq 0$. Hence, $a_0b_1 + a_1b_0 = 0$ shows that $b_0 = 0$. So we find that $b_0^2 = 0$, $b_0b_1 + b_1b_0 = 0$ and thus $(g(x))^2 \in \langle x^2 \rangle$. Hence, $\langle x^2 \rangle$ is a primary ideal. Now $\langle x^2 \rangle \subseteq \langle x \rangle$ and $f(x) \in \sqrt{\langle x^2 \rangle}$ if and only if $(f(x))^n \in \langle x^2 \rangle$ for some positive integer n. This is true if and only if the constant term of $f(x)$ is zero, i.e., if and only if $f(x) \in \langle x \rangle$.

Exercise 11 Show that a commutative ring R with 1 is isomorphic to a subdirect sum of a family of fields if and only if the intersection of all maximal ideals of R is $\{0\}$.

Solution: Suppose R is isomorphic to a subdirect sum of a family of fields $\{F_i \mid i \in I\}$. Then there exists a subring T of $\Pi_{i\in I}F_i$ such that $T = \oplus^s_{i\in I}F_i$ and $R \simeq T$. Let $\alpha : R \to T$ be an isomorphism. Then $\pi_i \circ \alpha : R \to F_i$ is an epimorphism for all $i \in I$, where π_i is the ith canonical projection. Proceeding as in the proof of Theorem 10.1.14, we can show that

$$\cap_{i\in I}A_i = \{0\},$$

where A_i =Ker $\pi_i \circ \alpha$ for all $i \in I$. Now $R/A_i \simeq F_i$. Since F_i is a field, A_i is a maximal ideal for all $i \in I$. If A is the intersection of all maximal ideals of R, then $A \subseteq \cap_{i\in I}A_i = \{0\}$. Hence, $A = \{0\}$. Conversely, suppose that $A = \{0\}$, where $A = \cap_{i\in J}\{M_i \mid M_i$ is a maximal ideal of $R\}$. By Theorem 10.1.14, R is monomorphic to the subdirect sum of a family of rings $\{R/M_i \mid i \in J\}$. Since each M_i is a maximal ideal, we find that R/M_i is a field.

Exercises

1. Find all maximal and prime ideals of $\mathbb{Z}_{10}$.
2. Prove that $I = \{(5n, m) \mid n, m \in \mathbb{Z}\}$ is a maximal ideal of $\mathbb{Z}\times\mathbb{Z}$.
3. Find all ideals and maximal ideals of $\mathbb{Z}_{p^k}$, where p is a prime and k is a positive integer.
4. Let $I = \{a_0 + a_1x + \cdots + a_nx^n \in \mathbb{Z}[x] \mid 3$ divides $a_0\}$. Show that I is a prime ideal of $\mathbb{Z}[x]$. Is I a maximal ideal?
5. Let I be an ideal of a ring R. Prove that the following conditions are equivalent.
   (i) I is a prime ideal.
   (ii) If $a, b \in R\backslash I$, then there exists $c \in R$ such that $acb \in R\backslash I$.
6. Let R be a finite commutative ring with 1. Show that in R, every prime ideal $I \neq R$ is a maximal ideal.
7. Let R be a Boolean ring. Prove that a nonzero proper ideal I of R is a prime ideal if and only if it is a maximal ideal.
8. Let R be a ring with 1. Prove that a nonzero proper ideal I of R is a maximal ideal if and only if the quotient ring R/I is a simple ring.
9. Let I be an ideal of a ring R. If P is a prime ideal of the quotient ring R/I, prove that there exists a prime ideal J of R such that $I \subseteq J$ and $J/I = P$.
10. Let R be a commutative ring with 1. Prove that there exists an epimorphism from R onto some field.
11. Let I be an ideal of a ring R with 1. Prove that the quotient ring R/I is a division ring if and only if I is a maximal right ideal.

12. For all $r \in \mathbb{R}$, show that $I_r = \{f(x) \in \mathbb{R}[x] \mid f(r) = 0\}$ is a maximal ideal of $\mathbb{R}[x]$ and $\mathbb{R}[x]/I_r \simeq \mathbb{R}$. Also, prove that $\cap_{r\in\mathbb{R}} I_r = \{0\}$.

13. Consider the polynomial ring $K[x]$ over a field K. Let $a \in K$. Define the mapping $\phi_a : K[x] \to K$ by $\phi_a(f(x)) = f(a)$ for all $f(x) \in K[x]$. Show that ϕ_a is an epimorphism and Ker ϕ_a is a maximal ideal of $K[x]$.

14. Let R be a PID.

    (i) Prove that every nonzero nonunit element is divisible by a prime element.

    (ii) If $\{I_n\}_{n\in\mathbb{N}}$ is a sequence of ideals of R such that $I_1 \subseteq I_2 \subseteq \cdots \subseteq I_n \subseteq \cdots$, prove that there exists a positive integer n such that $I_n = I_{n+1} = \cdots$.

    (iii) Prove that every nonzero nonunit can be expressed as a finite product of prime elements.

15. Let $\{I_\alpha\}$ be a collection of prime ideals in a commutative ring R such that $\{I_\alpha\}$ forms a chain. Prove that $\cap_\alpha I_\alpha$ and $\cup_\alpha\ I_\alpha$ are prime ideals of R.

16. If I_1 and I_2 are ideals of a commutative ring 1, prove that $\sqrt{I_1 \cap I_2} = \sqrt{I_1} \cap \sqrt{I_2}$.

17. Let R be a commutative ring with 1 and Q_i, $i = 1, 2, \ldots, n$, be ideals in R. Set $Q = \cap_{i=1}^n Q_i$. Prove that if $\sqrt{Q_i} = P$ for some ideal P of R, $i = 1, 2, \ldots, n$, then $\sqrt{Q} = P$. If $\sqrt{Q_i} = P$, $i = 1, 2, \ldots, n$, and each Q_i is primary, prove that Q is primary.

18. If I is an ideal of a commutative ring R with 1 such that $\sqrt{I}$ is a maximal ideal, prove that I is a primary ideal.

19. In the polynomial ring $\mathbb{Z}[x]$, prove the following.

    (i) $I = \{f(x) \in \mathbb{Z}[x]$ |the constant term of $f(x)$ is divisible by $4\}$ is a primary ideal with $J = \langle x, 2\rangle$ as its associated prime ideal.

    (ii) The ideal $\langle x, 6\rangle$ is not a primary ideal.

20. Prove that every prime ideal is a primary ideal in a commutative ring.

21. Let M be an ideal of a commutative ring R. Prove that R/M is a field if and only if M is a maximal ideal and $x^2 \in M$ implies $x \in M$ for all $x \in R$.

22. Prove that in a PID every nontrivial ideal I can be expressed as a finite product of prime ideals $I = P_1 \cdots P_n$ such that P_1, P_2, ..., P_n are determined uniquely up to order.

23. An ideal P of a ring R is called a **semiprime ideal** if for any ideal I of R, $I^2 \subseteq P$ implies that $I \subseteq P$.

    (i) Prove that an ideal P of R is a semiprime ideal if and only if the quotient ring R/P contains no nonzero nilpotent ideals.

    (ii) If R is a commutative ring with 1, prove that an ideal P of R is a semiprime ideal if and only if $\sqrt{P} = P$.

24. A commutative ring R with 1 is called a **local ring** if R has only one maximal ideal. Prove the following.

    (i) $\mathbb{Z}_8$ and $\mathbb{Z}_9$ are local rings.

    (ii) In a local ring, all nonunits form a maximal ideal.

    (iii) In a local ring R, for all $r, s \in R, r + s = 1$ implies either r is a unit or s is a unit.

25. Let p be a prime integer and $\mathbb{Q}_p = \{\frac{a}{b} \in \mathbb{Q} \mid p$ does not divide $b\}$. Show that $\mathbb{Q}_p$ is a local ring under the usual addition and multiplication of rational numbers.

26. Let R be a field and T be the set of all sequences $\{a_n\}$ of elements of R. Then $(T, +, \cdot)$ is a ring, where $+$ and $\cdot$ are defined as in Worked-Out Exercise 6 (page 193). Prove the following.

    (i) The set I of all nonunits of T is a maximal ideal of T.

    (ii) I is the only maximal ideal of T.

    (iii) T is a local ring.

27. Let $R = R_1 \oplus R_2 \oplus \cdots \oplus R_n$ be the direct sum of the finite family of rings $\{R_1, R_2, \ldots, R_n\}$, where each R_i contains an identity. Prove the following:

    (i) If M_i is a maximal ideal of R_i $(1 \le i \le n)$, then $R_1 \oplus R_2 \oplus \cdots \oplus R_{i-1} \oplus M_i \oplus R_{i+1} \oplus \cdots \oplus R_n$ is a maximal ideal of R.

    (ii) Every maximal ideal M of R is of the form

$$R_1 \oplus R_2 \oplus \cdots \oplus R_{i-1} \oplus M_i \oplus R_{i+1} \oplus \cdots \oplus R_n,$$

    where M_i is a maximal ideal of R_i for some i $(1 \le i \le n)$.

28. Show that the ring $\mathbb{Z}$ is isomorphic to a subdirect sum of a family of fields.

29. An ideal I of a ring R is called a **minimal ideal** if $I \neq \{0\}$ and there does not exist any ideal J of R such that $\{0\} \neq J \subset I$. If I is a minimal ideal of a commutative ring R with 1, prove that either $I^2 = \{0\}$ or $I = eR$ for some idempotent $e \in R$.

30. In the following exercises, write the proof if the statement is true; otherwise, give a counterexample.

    (i) Let R be a commutative ring with 1 and P be a prime ideal of R such that $P \neq R$. If the quotient ring R/P contains a finite number of elements, then R/P is a field.

    (ii) In a PID different from a field, there exists a prime element.

    (iii) In a PID, every proper prime ideal is a maximal prime ideal.

    (iv) The intersection of two prime ideals of a ring R is a prime ideal of R.

    (v) If I is a prime ideal of a ring R, then $I[x]$ is also a prime ideal of $R[x]$.

    (vi) If I is a maximal ideal of a ring R, then $I[x]$ is also a maximal ideal of $R[x]$.

    (vii) A commutative ring with 1 and with only a finite number of maximal ideals is a field.

    (viii) In the ring $\mathbb{Z}$, the ideal $\langle 5 \rangle$ is a maximal ideal, but in the ring $\mathbb{Z}[i]$, the ideal $\langle 5 \rangle$ is not a maximal ideal.

Chapter 15

Modules and Vector Spaces

15.1 Modules and Vector Spaces

Our main interest here is to set down only the results of vector spaces which are needed for our study of fields in the next chapter. We do this in such a way that the reader will become acquainted with the notion of a module.

Definition 15.1.1 *Let R be a ring. A commutative group $(M,+)$ is called a* ***left R-module*** *or a* ***left module*** *over R with respect to a mapping $\cdot : R \times M \to M$ if for all $r,s \in R$ and $m,m' \in M$,*
(i) $r \cdot (m+m') = r \cdot m + r \cdot m'$,
(ii) $r \cdot (s \cdot m) = (rs) \cdot m$,
(iii) $(r+s) \cdot m = r \cdot m + s \cdot m$.
If R has an identity 1 and if $1 \cdot m = m$ for all $m \in M$, then M is called a ***unitary*** *or* ***unital left R-module.***

A **right R-module** can be defined in a similar fashion.

In the above definition, we used the same notation for the addition in the ring R and the addition in the group M. We also used the same notation for the multiplication in R and the multiplication between the elements of R and M. It should be clear to the reader by now that there are actually four distinct operations involved. We write rm for $r \cdot m$.

Example 15.1.2 *In a ring R, every left ideal is a left R-module and every right ideal is a right R-module. In particular, R is a left and right R-module.*

Example 15.1.3 *Every commutative group M is a module over the ring of integers $\mathbb{Z}$. For $n \in \mathbb{Z}$ and $a \in M$, the element na is defined to be a added to itself n times if n is positive and $-a$ added to itself $|n|$ times if n is negative. $0a$ is defined to be the zero element of M. Under these definitions, M becomes a unitary left $\mathbb{Z}$-module.*

Let M be any commutative group and R be any ring. If we define $rm = 0$ for all $r \in R$, $m \in M$, then M forms a left R-module, called a trivial module.

Since all results that are true for left R-modules are also true for right R-modules, we prove results only for left R-modules. From now on, unless stated otherwise, by an R-module, we mean a left R-module.

Definition 15.1.4 *Let M be an R-module and N be a nonempty subset of M. Then N is called a* ***submodule*** *of M if N is a subgroup of M and for all $r \in R$, $a \in N$, we have $ra \in N$.*

It is clear that a submodule of an R-module is itself an R-module.

Using arguments similar to those used for subgroups and ideals, one can show that the intersection of any nonempty collection of submodules of an R-module is again a submodule.

Definition 15.1.5 *Let X be a subset of an R-module M. Then the submodule of M* ***generated*** *by X is defined to be the intersection of all submodules of M which contain X and is denoted by $\langle X \rangle$. X is called a* ***basis*** *of $\langle X \rangle$ if no proper subset of X generates $\langle X \rangle$. If $M = \langle X \rangle$ and X is a finite set, then M is said to be* ***finitely generated****. When $X = \{x\}$ and $M = \langle \{x\} \rangle$, then M is called a* ***cyclic*** *R-module and in this case we write $M = \langle x \rangle$.*

We ask the reader to prove that any finitely generated module has a finite basis.

The proof of the following theorem is similar to that of the corresponding theorem for ideals, Theorem 8.2.9. Hence, we omit its proof.

Theorem 15.1.6 *Let M be an R-module and X be a nonempty subset of M. Then*

$$\begin{aligned}\langle X\rangle &= \{\textstyle\sum_{i=1}^{k} r_i x_i + \sum_{j=1}^{l} n_j x'_j \mid r_i \in R, n_j \in \mathbb{Z}, x_i, x'_j \in X,\\ &\quad 1 \le i \le k, 1 \le j \le l, k, l \in \mathbb{N}\}.\end{aligned}$$

If M is a unitary R-module, then

$$\langle X\rangle = \{\sum_{i=1}^{k} r_i x_i \mid r_i \in R,\ x_i \in X, 1 \le i \le k,\ k \in \mathbb{N}\}. \blacksquare$$

Example 15.1.7 *(i) $\mathbb{Q}$ is a $\mathbb{Q}$-module. If N is a submodule of $\mathbb{Q}$, then N is a left ideal of $\mathbb{Q}$. Since $\mathbb{Q}$ is a field, the only left ideals of $\mathbb{Q}$ are $\{0\}$ and $\mathbb{Q}$. Hence the submodules of $\mathbb{Q}$ are $\{0\}$ and $\mathbb{Q}$.*

(ii) We know that $\mathbb{Q}\oplus\mathbb{Q}$ is a commutative group. For all $x \in \mathbb{Q}$ and for all $(a,b) \in \mathbb{Q}\oplus\mathbb{Q}$, define $x(a,b) = (xa, xb)$. Then $\mathbb{Q}\oplus\mathbb{Q}$ is a $\mathbb{Q}$-module. We now determine all submodules of $\mathbb{Q}\oplus\mathbb{Q}$. Let M be a nonzero $\mathbb{Q}$-submodule of $\mathbb{Q}\oplus\mathbb{Q}$.

Case 1: Suppose for all $(a,b) \in M$, $b = 0$. Now there exists $(a,0) \in M$ such that $a \neq 0$. Then $(1,0) = \frac{1}{a}(a,0) \in M$. Thus, $M = \mathbb{Q}\oplus\{0\}$.

Case 2: Suppose for all $(a,b) \in M$, $a = 0$. Now there exists $(0,b) \in M$ such that $b \neq 0$. Then $(0,1) = \frac{1}{b}(0,b) \in M$. Thus, $M = \{0\}\oplus\mathbb{Q}$.

Case 3: Suppose there exists $(a,b) \in M$ such that $a \neq 0$, $b \neq 0$.

Case 3a: Suppose $M = \langle(a,b)\rangle$. Then M is a cyclic submodule of $\mathbb{Q}\oplus\mathbb{Q}$ generated by (a,b).

Case 3b: Suppose $M \neq \langle(a,b)\rangle$. Then $\langle(a,b)\rangle \subset M$. Thus, there exists $(a',b') \in M\setminus\langle(a,b)\rangle$. Then $a' \neq 0$ or $b' \neq 0$. Suppose that $a' = 0$. Then $(0,1) = \frac{1}{b'}(0,b') \in M$. Therefore, $(a,0) = (a,b) - (0,1)b \in M$. Hence, $(1,0) = \frac{1}{a}(a,0) \in M$. Thus, $(1,0), (0,1) \in M$. This implies that $M = \mathbb{Q}\oplus\mathbb{Q}$. Similarly, if $b' = 0$, then $M = \mathbb{Q}\oplus\mathbb{Q}$.

Now suppose that $a' \neq 0$ and $b' \neq 0$. If $\frac{a}{a'} = \frac{b}{b'} = t$ (say), then $t(a',b') = (ta', tb') = (\frac{a}{a'}a', \frac{b}{b'}b') = (a,b) \in \langle(a,b)\rangle$, which is a contradiction. Therefore, $\frac{a}{a'} \neq \frac{b}{b'}$ and so $ab' - ba' \neq 0$. Let $(p,q) \in \mathbb{Q}\oplus\mathbb{Q}$. Choose $t = \frac{pb'-qa'}{ab'-ba'}$ and $s = \frac{qa-pb}{ab'-ba'}$. Then $(p,q) = t(a,b) + s(a',b') \in M$. Thus, $\mathbb{Q}\oplus\mathbb{Q} \subseteq M$. Hence, $M = \mathbb{Q}\oplus\mathbb{Q}$.

Consequently, if M is a $\mathbb{Q}$-submodule of $\mathbb{Q}\oplus\mathbb{Q}$, then M is of the following form:

(i) $M = \{0\}$, or

(ii) $M = \{0\}\oplus\mathbb{Q} = \langle(0,1)\rangle$, or

(iii) $M = \mathbb{Q}\oplus\{0\} = \langle(1,0)\rangle$, or

(iv) $M = \langle(a,b)\rangle$, $a \neq 0$, $b \neq 0$, $a, b \in \mathbb{Q}$, or

(v) $M = \mathbb{Q}\oplus\mathbb{Q}$.

This also proves that M is finitely generated.

Definition 15.1.8 *Let F be a field. A unitary (left) F-module M is called a (**left**) **vector space** over F. The elements of M are called **vectors** and the elements of F are called **scalars**. A submodule of M is called a **subspace** of M. If X is a subset of M such that $M = \langle X\rangle$, then X is said to **span** or **generate** M and M is called the **span** of X over F.*

Example 15.1.9 *Let F be any field and F^n denote the Cartesian product of F with itself $n$times. Then F^n becomes a vector space over F under the following definitions: For all $(a_1, a_2, \ldots, a_n)$, $(b_1, b_2, \ldots, b_n) \in F^n$ and $a \in F$*

$$\begin{aligned}(a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) &= (a_1 + b_1, +a_2 + b_2, \ldots, a_n + b_n),\\ a(a_1, a_2, \ldots, a_n) &= (aa_1, aa_2, \ldots, aa_n).\end{aligned}$$

The set

$$X = \{(1,0,0,\ldots,0), (0,1,0,\ldots,0), \ldots, (0,0,0,\ldots,1)\}$$

spans F^n since for all $(a_1, a_2, \ldots, a_n) \in F^n$,

$$(a_1, a_2, \ldots, a_n) = a_1(1,0,0,\ldots,0) + a_2(0,1,0,\ldots,0) + \cdots + a_n(0,0,0,\ldots,1).$$

When $n = 2$ or 3 and F is the field of real numbers, then the vector space F^n over F is the one usually encountered in elementary analytical geometry.

By Example 15.1.9, $\mathbb{R}^3$ is a vector space over $\mathbb{R}$.

Example 15.1.10 *Consider the vector space $\mathbb{R}^3$ over $\mathbb{R}$. Let*

$$U = \{(a, b, c) \in \mathbb{R}^3 \mid 2a + 3b + 5c = 0\}.$$

Then U is a subspace of $V_3(\mathbb{R})$. Let

$$U_1 = \{(a, b, c) \in \mathbb{R}^3 \mid 2a + 3b + 5c = 5\}.$$

Now $(0, 0, 1)$ and $(1, 1, 0) \in U_1$, but $(0, 0, 1) + (1, 1, 0) \notin U_1$. Hence, U_1 is not a subspace of $\mathbb{R}^3$.

Example 15.1.11 *Let V be a vector space over F. Then $\{0\}$ and V are subspaces of V. These are called **trivial subspaces** of V.*

Theorem 15.1.12 *Let V be a vector space over F and S be a nonempty subset of V. Then S is a subspace of V if and only if for all $a \in F$ and for all $x, y \in S$, $ax + y \in S$.*

Proof. Suppose S is a subspace of V. Then for all $a \in F$ and for all $x, y \in S$, $ax \in S$ and so $ax + y \in S$. Conversely, suppose for all $a \in F$ and for all $x, y \in S$, $ax + y \in S$. Since $S \neq \emptyset$, there exists $x \in S$. By Exercise 2 (page 226), $-x = (-1)x$. Therefore, $0 = -x + x = (-1)x + x \in S$. Hence, for all $x \in S$, $-x = (-1)x + 0 \in S$. Also, for all $x, y \in S$, $x + y = 1x + y \in S$. S inherits the associative and commutative laws. Thus, $(S, +)$ is a commutative group. Now for all $a \in F$ and for all $x \in S$, $ax = ax + 0 \in S$. Therefore, S is a vector space over F since the other properties are inherited. ■

Theorem 15.1.13 *Let V be a vector space over F and $\{U_\alpha \mid \alpha \in I\}$ be any nonempty collection of subspaces of V. Then $\cap_{\alpha \in I} U_\alpha$ is a subspace of V.*

Proof. First note that $0 \in U_\alpha$ for all $\alpha \in I$ and so $0 \in \cap_{\alpha \in I} U_\alpha$. Therefore, $\cap_{\alpha \in I} U_\alpha \neq \emptyset$. Let $a \in F$ and $x, y \in \cap_{\alpha \in I} U_\alpha$. Then $x, y \in U_\alpha$ for all α. Since U_α is a subspace of V, $ax + y \in U_\alpha$ for all $\alpha \in I$ and so $ax + y \in \cap_{\alpha \in I} U_\alpha$. Thus, $\cap_{\alpha \in I} U_\alpha$ is a subspace of V by Theorem 15.1.12. ■

Theorem 15.1.14 *Let V be a vector space over F and S be a nonempty subset of V. Then*

$$\langle S \rangle = \{\sum a_i s_i \mid a_i \in F,\ s_i \in S\},$$

where $\sum a_i s_i$ is a finite sum.

Proof. Let $U = \{\sum a_i s_i \mid a_i \in F,\ s_i \in S\}$. Let $a \in F$ and $\sum a_i s_i,\ \sum b_j s_j \in U$. Then $a(\sum a_i s_i) + \sum b_j s_j = \sum (aa_i) s_i + \sum b_j s_j \in U$ and so U is a subspace of V by Theorem 15.1.12. Since for all $s \in S$, $s = 1s \in U$, $U \supseteq S$. Thus, $U \supseteq \langle S \rangle$ since $\langle S \rangle$ is the smallest subspace of V containing S. Let $\sum a_i s_i \in U$. Then since $s_i \in S \subseteq \langle S \rangle$, $a_i s_i \in \langle S \rangle$. Thus, $\sum a_i s_i \in \langle S \rangle$, whence $U \subseteq \langle S \rangle$. ■

Definition 15.1.15 *Let V be a vector space over the field F. A subset X of V is called **linearly independent** over F if for every finite number of distinct elements $x_1, x_2, \ldots, x_n \in X$, $a_1 x_1 + a_2 x_2 + \cdots + a_n x_n = 0$ implies that $a_1 = a_2 = \cdots a_n = 0$ for any finite set of scalars $\{a_1, a_2, \ldots, a_n\}$. Otherwise X is called **linearly dependent** over F.*

The set X in Example 15.1.9 is linearly independent over F. $\{0\}$ is linearly dependent over F.

Definition 15.1.16 *Let V be a vector space over F. A subset A of V is called a **basis** for V over F if A spans V, i.e., $V = \langle A \rangle$, and A is linearly independent over F.*

Consider the zero vector space, $\{0\}$, over the field F. We note that the empty subset, $\emptyset$, is linearly independent over F vacuously and that $\emptyset$ spans $\{0\}$. Hence, $\emptyset$ is a basis for $\{0\}$.

Example 15.1.17 *The set*

$$X = \{(1, 0, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, 0, 0, \ldots, 1)\}$$

of Example 15.1.9 *is a basis for F^n. We showed there that X spans F^n over F. Suppose*

$$(0, 0, \ldots, 0) = a_1(1, 0, 0, \ldots, 0) + a_2(0, 1, 0, \ldots, 0) + \cdots + a_n(0, 0, 0, \ldots, 1).$$

Then $(0, 0, \ldots, 0) = (a_1, a_2, \ldots, a_n)$. Therefore, we must have $a_i = 0$ for $i = 1, 2, \ldots, n$. Thus, X is linearly independent.

Theorem 15.1.18 *Let V be a vector space over F and S be a subset of V. If $s \in \langle S \rangle$, then $\langle S \cup \{s\} \rangle = \langle S \rangle$.*

Proof. Clearly $\langle S \rangle \subseteq \langle S \cup \{s\} \rangle$. If $S = \emptyset$, then $\langle S \rangle = \{0\}$ and so $s = 0$. Hence, $\langle S \cup \{s\} \rangle = \langle \{0\} \rangle = \{0\} = \langle S \rangle$. Suppose $S \neq \emptyset$. Let $\sum a_i s_i + as \in \langle S \cup \{s\} \rangle$, where $s_i \in S$. Then $\sum a_i s_i, as \in \langle S \rangle$ and so $\sum a_i s_i + as \in \langle S \rangle$. Hence, $\langle S \cup \{s\} \rangle = \langle S \rangle$. ■

Theorem 15.1.19 *Let V be a vector space over F and $A = \{x_1, x_2, \ldots, x_r\}$ be a subset of V which spans V. Let B be any linearly independent set of vectors in V. Then B contains at most r vectors.*

Proof. If B contains less than r vectors, the theorem is true. Suppose B contains at least r vectors, say, $y_1, y_2, \ldots, y_r \in B$. Then since A spans V,

$$y_1 = \sum_{i=1}^{r} a_{i1} x_i$$

and since $y_1 \neq 0$, not all $a_{i1} = 0$, say, $a_{11} \neq 0$. Thus,

$$x_1 = \sum_{i=2}^{r} (-a_{11}^{-1} a_{i1}) x_i + a_{11}^{-1} y_1.$$

This implies that $x_1 \in \langle \{y_1, x_2, \ldots, x_r\} \rangle$. Hence, $\langle \{y_1, x_2, \ldots, x_r\} \rangle = V$ by Theorem 15.1.18. Assume $\langle \{y_1, y_2, \ldots, y_k, x_{k+1}, \ldots, x_r\} \rangle = V$, the induction hypothesis. Then

$$y_{k+1} \in \langle \{y_1, y_2, \ldots, y_k, x_{k+1}, \ldots, x_r\} \rangle.$$

Thus,

$$y_{k+1} = \sum_{i=1}^{k} a_{i,k+1} y_i + \sum_{i=k+1}^{r} a_{i,k+1} x_i$$

and not all $a_{i,k+1} = 0$ for $i = k+1, \ldots, r$, say, $a_{k+1,k+1} \neq 0$. This implies that

$$x_{k+1} = \sum_{i=1}^{k} (-a_{k+1,k+1}^{-1} a_{i,k+1}) y_i + \sum_{i=k+2}^{r} (-a_{k+1,k+1}^{-1} a_{i,k+1}) x_i + a_{k+1,k+1}^{-1} y_{k+1}.$$

Thus, $x_{k+1} \in \langle \{y_1, y_2, \ldots, y_k, y_{k+1}, x_{k+2}, \ldots, x_r\} \rangle$. Hence,

$$V = \langle \{y_1, y_2, \ldots, y_k, y_{k+1}, x_{k+2}, \ldots, x_r\} \rangle$$

by Theorem 15.1.18. Thus, $\langle \{y_1, y_2, \ldots, y_r\} \rangle = V$ by induction. If there exists $y \in B$ such that $y \neq y_i$, $i = 1, 2, \ldots, r$, then $y = \sum_{i=1}^{r} a_i y_i$ and so $0 = \sum_{i=1}^{r} a_i y_i + (-1)y$ and since $-1 \neq 0$, $y_1, y_2, \ldots, y_r, y$ are not linearly independent, a contradiction. Therefore, y does not exist and so $B = \{y_1, y_2, \ldots, y_r\}$. ■

Theorem 15.1.20 *Let V be a vector space over F, $A = \{x_1, \ldots, x_r\}$, and $B = \{y_1, \ldots, y_s\}$ be two bases for V. Then $r = s$.*

Proof. Since A spans V and B is linearly independent, $s \leq r$ by Theorem 15.1.19. Similarly, $r \leq s$. ■

Definition 15.1.21 *Let V be a vector space over F. If V is spanned by a finite set of vectors, then V is called* ***finite dimensional*** *over F.*

Lemma 15.1.22 *Let V be a vector space over F and A be a linearly independent subset of V. If $x \in V$ and $x \notin \langle A \rangle$, then $A \cup \{x\}$ is linearly independent.*

Proof. Let $x_1, \ldots, x_n \in A$. Suppose $0 = a_1 x_1 + a_2 x_2 + \cdots + a_r x_r + ax$. Suppose $a \neq 0$. Then

$$x = (-a)^{-1} a_1 x_1 + \cdots + (-a)^{-1} a_r x_r \in \langle A \rangle,$$

a contradiction. Thus, $a = 0$. Hence, $0 = a_1 x_1 + a_2 x_2 + \cdots + a_r x_r$. Since $\{x_1, x_2, \ldots, x_r\}$ is linearly independent, $a_1 = 0, \ldots, a_r = 0$. Thus, $A \cup \{x\}$ is linearly independent. ■

Theorem 15.1.23 *Let V be a finite dimensional vector space over F. Then V has a basis.*

Proof. If $V = \{0\}$, then $\emptyset$ is a basis for V. We now assume that $V \neq \{0\}$. Let $x_1 \in V$ be such that $x_1 \neq 0$. Then x_1 is linearly independent. If $\langle x_1 \rangle \neq V$, then there exists $x_2 \in V$ such that $x_2 \notin \langle x_1 \rangle$. By Lemma 15.1.22, x_1 and x_2 are linearly independent. Suppose $x_1, \ldots, x_k \in V$ are linearly independent and $\langle \{x_1, \ldots, x_k\} \rangle \neq V$. Then there exists $x_{k+1} \in V$ such that $x_{k+1} \notin \langle \{x_1, \ldots, x_k\} \rangle$. Therefore, $x_1, \ldots, x_k$, x_{k+1} are linearly independent. Since V is finite dimensional, V is spanned by, say, r vectors. By Theorem 15.1.19, any linearly independent set of vectors in V cannot have more than r vectors. Hence, if we continue the above process of constructing x_i's, then there must exist a positive integer s such that $\{x_1, \ldots, x_s\}$ is linearly independent, $\langle \{x_1, \ldots, x_s\} \rangle = V$, and $s \leq r$. Thus, $\{x_1, \ldots, x_s\}$ is a basis of V.
■

Theorem 15.1.23 gives us a method for constructing a basis for a finite dimensional vector space V of dimension n over F. We first take any nonzero vector x_1 of V. If $\langle x_1 \rangle = V$, then $\{x_1\}$ is a basis of V. If $\langle x_1 \rangle \subset V$, then we take any $x_2 \in V$, $x_2 \notin \langle x_1 \rangle$. Then by Lemma 15.1.22 $\{x_1, x_2\}$ is linearly independent over F. If $\langle \{x_1, x_2\} \rangle = V$, then $\{x_1, x_2\}$ is a basis for V over F. If $\langle \{x_1, x_2\} \rangle \subset V$, we can choose $x_3 \in V, x_3 \notin \langle \{x_1, x_2\} \rangle$ and so on. In a finite number of steps, precisely n steps, we must arrive at a basis for V over F.

Definition 15.1.24 *Let V be a finite dimensional vector space over F. The **dimension** V is the number of elements in a basis for V.*

From the statements following Definition 15.1.16, it follows that the zero vector space, $\{0\}$, is of dimension 0.

Theorem 15.1.25 *Let V be a vector space of dimension n over the field F. Then $X = \{x_1, x_2, \ldots, x_n\}$ is a basis of V if and only if every vector in V is a unique linear combination of $x_1, x_2, \ldots, x_n$ over F.*

Proof. Suppose X is a basis of V over F. Then by Theorem 15.1.14, every vector $v \in V$ is a linear combination of $x_1, x_2, \ldots, x_n$. Let

$$v = a_1x_1 + \cdots + a_nx_n = b_1x_1 + \cdots + b_nx_n$$

be any two linear combinations of $x_1, x_2, \ldots, x_n$. Then

$$0 = (a_1 - b_1)x_1 + \cdots + (a_n - b_n)x_n.$$

The linear independence of X over F implies that $a_1 - b_1 = 0, \ldots, a_n - b_n = 0$. That is, the representation of v as a linear combination of $x_1, x_2, \ldots, x_n$ is unique. Conversely, suppose every vector in V is a unique linear combination of $x_1, x_2, \ldots, x_n$ over F. Then clearly X generates V over F. Suppose $0 = a_1x_1 + \cdots + a_nx_n$ for $a_i \in F$. Since also $0 = 0x_1 + \cdots + 0x_n$, we have $a_i = 0$, $i = 1, \ldots, n$. Thus, X is linearly independent over F. By definition, X is a basis of V over F. ■

We now show that every nonzero vector space, not necessarily finite dimensional, has a basis. For this we prove the following lemma.

Lemma 15.1.26 *Let V be a vector space over a field F and X be a nonempty subset of V. Then X is a basis for V if and only if X is a maximal linearly independent set over F.*

Proof. If X is a basis for V, then X is linearly independent over F and $\langle X \rangle = V$. Let $y \in V$, $y \notin X$. Then $V = \langle X \rangle \subseteq \langle X \cup \{y\} \rangle \subseteq V$ so that $V = \langle X \cup \{y\} \rangle$. Since the proper subset X of $X \cup \{y\}$ also generates V, $X \cup \{y\}$ cannot be linearly independent over F. Thus, X is a maximal linearly independent set over F. Conversely, let X be a maximal linearly independent set over F. It suffices to show that $V = \langle X \rangle$. If $V \supset \langle X \rangle$, then there exists $y \in V$, $y \notin \langle X \rangle$. By Lemma 15.1.22, $X \cup \{y\}$ is linearly independent over F, which contradicts the maximality of X. Thus, $V = \langle X \rangle$. ■

Theorem 15.1.27 *Let V be a vector space over the field F. Then V has a basis.*

Proof. If $V = \{0\}$, then $\emptyset$ is a basis for V. We now assume that $V \neq \{0\}$. Let x be a nonzero element of V. Then $\{x\}$ is a linearly independent subset of V. Let T be the set of all linearly independent subsets of V that contain $\{x\}$. Clearly $T \neq \emptyset$. T is a poset with respect to the set inclusion relation. By Zorn's lemma, we can show that T has a maximal element, say, X. Then X is a maximal linearly independent subset of V and by Lemma 15.1.26, it follows that X is a basis of V. ■

Finally, we state the following theorem without proof. The finite dimensional case was proved in Theorem 15.1.20.

Theorem 15.1.28 *Let V be a vector space over a field F. If A and B are two bases of V, then $|A| = |B|$.* ■

From Theorem 15.1.27, we find that a vector space V over a field F has a basis B. If B is a basis for V over F, then $|B|$ is called the **dimension** of V over F.

Worked-Out Exercises

♢ **Exercise 1** Let V be a vector space of dimension n. Show that any set of n linearly independent vectors is a basis of V.

Solution: Let B be a set of n linearly independent vectors. Suppose $V \neq \langle B \rangle$. Let $y \in V$ be such that $y \notin \langle B \rangle$. Then $B \cup \{y\}$ is a set of $n+1$ linearly independent vectors by Lemma 15.1.22, a contradiction to Theorem 15.1.19. Hence, B is a basis of V.

♢ **Exercise 2** Let $u_1 = (0,1,1,0)$, $u_2 = (1,0,1,0)$, and $u_3 = (-1,-2,0,0)$ be three vectors in $\mathbb{R}^4$. Show that $\{u_1, u_2, u_3\}$ is a linearly independent set. Extend this set to a basis of $\mathbb{R}^4$.

Solution: Let $a_1, a_2, a_3 \in \mathbb{R}$ be such that

$$a_1u_1 + a_2u_2 + a_3u_3 = 0.$$

Then $a_2 - a_3 = 0$, $a_1 - 2a_3 = 0$, and $a_1 + a_2 = 0$. From this, it follows that $a_1 = a_2 = a_3 = 0$. Hence, $\{u_1, u_2, u_3\}$ is a linearly independent set. Suppose

$$(0,0,0,1) \in \langle \{u_1, u_2, u_3\} \rangle .$$

Then there exists $b_1, b_2, b_3 \in \mathbb{R}$ such that

$$b_1u_1 + b_2u_2 + b_3u_3 = (0,0,0,1).$$

Thus, $b_2 - b_3 = 0$, $b_1 - 2b_3 = 0$, $b_1 + b_2 = 0$, and $1 = 0$, a contradiction. Therefore, $e_4 = (0,0,0,1) \notin \langle \{u_1, u_2, u_3\} \rangle$. Hence, $\{u_1, u_2, u_3, e_4\}$ is a linearly independent set of vectors in $\mathbb{R}^4$. Since the dimension of $\mathbb{R}^4$ is 4, $\{u_1, u_2, u_3, e_4\}$ is a basis.

♢ **Exercise 3** Let V be a nonzero vector space of dimension n. Let X be a finite subset of V such that $V = \langle X \rangle$. Show that X contains a subset Y such that Y is a basis of V.

Solution: Let $X = \{x_1, x_2, \ldots, x_t\}$. Clearly $t \geq n$. Since $V \neq \{0\}$, X contains a nonzero element. Thus, X contains a linearly independent subset. If X is linearly independent, then X is a basis of V and $n = t$. Suppose X is not linearly independent. Then there exists x_i, say, x_t, such that $x_t \in \langle \{x_1, x_2, \ldots, x_{t-1}\} \rangle$. Then $V = \langle \{x_1, x_2, \ldots, x_{t-1}\rangle\}$. Let $s = t - n - 1$. By repeating the process finitely many times, we can show that there are s vectors $x_{i_1}, \ldots, x_{i_s} \in \{x_1,\ x_2,\ \ldots,\ x_{t-1}\}$ such that

$$x_{i_1}, \ldots, x_{i_s} \in \langle \{x_1, x_2, \ldots, x_{t-1}\} \backslash \{x_{i_1}, \ldots, x_{i_s}\} \rangle .$$

Let

$$Y = \{x_1, x_2, \ldots, x_{t-1}\} \backslash \{x_{i_1}, \ldots, x_{i_s}\}.$$

Then $Y \subseteq X$, $|Y| = n$, and $V = \langle Y \rangle$. If Y is not linearly independent, then there exists $y \in Y$ such that $y \in \langle Y \backslash \{y\} \rangle$. Then $V = \langle Y \backslash \{y\} \rangle$ and $|Y \backslash \{y\}| = n - 1$, a contradiction to the fact that the dimension of V is n.

Exercise 4 Let $T = \{(x,y,z) \in \mathbb{R}^3 \mid 2x + 3y + z = 0\}$. Show that T is a subspace of $V_3(\mathbb{R})$. Find a basis for T.

Solution: Since $(0,0,0) \in T$, $T \neq \emptyset$. Let $(x_1,\ y_1,\ z_1)$, $(x_2,\ y_2,\ z_2) \in T$ and $r \in \mathbb{R}$. Then $2x_1 + 3y_1 + z_1 = 0$ and $2x_2 + 3y_2 + z_2 = 0$. Hence, $2(x_1 + x_2) + 3(y_1 + y_2) + (z_1 + z_2) = 0$ and $2rx_1 + 3ry_1 + rz_1 = r(2x_1 + 3y_1 + z_1) = 0$. Therefore, $(x_1,\ y_1,\ z_1) + (x_2,\ y_2,\ z_2) \in T$ and $r(x_1,\ y_1,\ z_1) \in T$. Thus, T is a subspace of $V_3(\mathbb{R})$. Now $2x_1 + 3y_1 + z_1 = 0$ implies that $(x_1,\ y_1,\ z_1) = (x_1,\ y_1, -2x_1 - 3y_1) = x_1(1,0,-2) + y_1(0,1,-3)$. Since $(1,0,-2)$, $(0,1,-3) \in T$ and $(x_1,\ y_1,\ z_1)$ is an arbitrary element of T, $T = \langle \{(1,0,-2), (0,1,-3)\} \rangle$. It is easy to verify that $\{(1,0,-2),\ (0,1,-3)\}$ is a linearly independent set. Hence, $\{(1,0,-2),\ (0,1,-3)\}$ is a basis of T.

Exercises

1. For the vector space $\mathbb{R}^3$ over $\mathbb{R}$, determine whether or not the sets listed are bases of $\mathbb{R}^3$.
   (i) $\{(1,1,0), (1,1,1), (1,0,0)\}$.
   (ii) $\{(2,0,0), (0,2,0), (0,0,2)\}$.
   (iii) $\{(-1,0,0), (0,-1,0), (0,0,-1)\}$.
   (iv) $\{(1,0,0), (1,1,0), (1,1,1), (0,1,0)\}$.
2. Let M be an R-module, $m \in M$ and $r \in R$. Prove that $r0 = 0$, $0m = 0$, and $-(rm) = (-r)m = r(-m)$.
3. Show that the intersection of two submodules of an R-module M is a submodule.

4. Show that the $\mathbb{Z}$-module $\mathbb{Q}$ has no finite set of generators.
5. Find all subspaces of the real vector space $\mathbb{R}^2$. Is is true that for any elements $u = (a, b)$ and $v = (c, d)$ of $\mathbb{R}^2$, there exists a nontrivial subspace W of $\mathbb{R}^2$ such that $u, v \in W$?
6. Let A, B, and C be submodules of an R-module M.
   (i) Prove that $A + B = \{a + b \mid a \in A,\ b \in B\}$ is a submodule of M.
   (ii) If $A \subseteq C$, prove that $A + (B \cap C) = (A + B) \cap C$.
7. Let M be an R-module and $a \in M$. Show that $T = \{ra + na \mid r \in R,\ n \in \mathbb{Z}\}$ is a submodule of M.
8. Let M be a unitary R-module. M is called a **simple** R-module if $M \neq \{0\}$ and the only submodules of M are M and $\{0\}$. Prove that M is simple if and only if M is generated by any nonzero element of M.
9. Let N be a submodule of a unitary R-module M and $a \in M$. Let
$$a + N = \{a + b \mid b \in N\}.$$
   Prove the following.
   (i) $a \in a + N$.
   (ii) For all $a, b \in M$, $a + N = b + N$ if and only if $a - b \in N$.
   (iii) For all $a, b \in M$, either $(a + N) \cap (b + N) = \emptyset$ or $a + N = b + N$.
10. Let N be a submodule of an R-module M. Let
$$M/N = \{a + N \mid a \in N\}.$$
    Define the following operations on M/N
$$\begin{aligned} (a + N) + (b + N) &= (a + b) + N \\ r(a + N) &= ra + N \end{aligned}$$
    for all $a + N, b + N \in M/N$, $r \in R$. Prove that M/N is an R-module.
11. Let V be a finite dimensional vector space over F. If U and W are two subspaces of V, prove the following:
    (i) $U + W = \{u + w \mid u \in U, w \in W\}$ is a subspace of V.
    (ii) $\dim U + \dim W = \dim(U + W) - \dim(U \cap W)$.
12. Let N be a submodule of an R-module M. N is called a **direct summand** of M if there exists a submodule P of M such that $M = N + P$ and $N \cap P = \{0\}$. In a finite dimensional vector space V over F, show that every subspace is a direct summand of V.
13. Write the proof if the statement is true; otherwise give a counterexample.
    (i) If $\{u, v, w\}$ is a linearly independent subset of a vector space V, then $\{u,\ u + v,\ u + v + w\}$ is also a linearly independent subset.
    (ii) If W is a subspace of a finite dimensional vector space V such that $\dim W = \dim V$, then $W = V$.
    (iii) Let V be a vector space over a field F. If $0 \neq v \in V$, then there exists a basis containing v.
    (iv) If S and T are two basis of a vector space V, then $S \cup T$ is a basis of V.

Chapter 16

Field Extensions

In this chapter, we study a special type of ring called a field. Results about fields have applications in number theory and the theory of equations. The theory of equations deals with roots of polynomials. It is here that our main interest lies. This interest leads us to an introduction of Galois theory.

The importance of the concept of a field was first recognized by Abel and Galois in their research on the solution of equations by radicals. However, the formal definition of a field appeared more than 70 years later. The works of Dedekind and Kronecker seem to be responsible for the entrance of the concept of a field into mathematics. However, in 1910, in his paper, *Algebraic Theorie der Köperer*, Steinitz gave the first abstract definition of a field. His work freed the concept of a field from the context of complex numbers.

16.1 Algebraic Extensions

Let us recall that the characteristic of a field F is either 0 or a prime p. By Theorem 8.1.9, the intersection of any collection of subfields of a field F is again a subfield of F. Hence, a field contains a subfield which has no proper subfield, namely, the intersection of all its subfields.

Definition 16.1.1 *A field F is called a **prime field** if F has no proper subfield.*

Theorem 16.1.2 *Let F be a field.*
(i) If the characteristic of F is 0, then F contains a subfield K such that $K \simeq \mathbb{Q}$.
(ii) If the characteristic of F is $p > 0$, then F contains a subfield K such that $K \simeq \mathbb{Z}_p$.

Proof. Define $f : \mathbb{Z} \to F$ by

$$f(n) = n1$$

for all $n \in \mathbb{Z}$, where 1 denotes the identity of F. Then f is a homomorphism.

(i) Suppose the characteristic of F is 0. Then Ker $f = \{0\}$ and so f is one-one. Define $f^* : \mathbb{Q} \to F$ by

$$f^*(\frac{a}{b}) = f(a)f(b)^{-1}$$

for all $\frac{a}{b} \in \mathbb{Q}$. Let $\frac{a}{b}, \frac{c}{d} \in \mathbb{Q}$. Now $\frac{a}{b} = \frac{c}{d}$ if and only if $ad = bc$ if and only if $f(ad) = f(bc)$ if and only if $f(a)f(d) = f(c)f(b)$ if and only if $f(a)f(b)^{-1} = f(c)f(d)^{-1}$ if and only if $f^*(\frac{a}{b}) = f^*(\frac{c}{d})$. Hence, f^* is a one-one function. Now

$$\begin{aligned} f^*(\tfrac{a}{b} + \tfrac{c}{d}) &= f^*(\tfrac{ad+bc}{bd}) \\ &= f(ad+bc)f(bd)^{-1} \\ &= (f(a)f(d) + f(b)f(c))f(b)^{-1}f(d)^{-1} \\ &= f(a)f(b)^{-1} + f(c)f(d)^{-1} \\ &= f^*(\tfrac{a}{b}) + f^*(\tfrac{c}{d}). \end{aligned}$$

Also,

$$\begin{aligned} f^*(\tfrac{a}{b} \cdot \tfrac{c}{d}) &= f^*(\tfrac{ac}{bd}) \\ &= f(ac)f(bd)^{-1} \\ &= f(a)f(c)f(b)^{-1}f(d)^{-1} \\ &= f(a)f(b)^{-1}f(c)f(d)^{-1} \\ &= f^*(\tfrac{a}{b})f^*(\tfrac{c}{d}). \end{aligned}$$

Thus, f^* is a homomorphism. Hence, $\mathbb{Q} \simeq \mathcal{I}(f^*)$, where $\mathcal{I}(f^*)$ is the image of f^*. Let $K = \mathcal{I}(f^*)$.

(ii) Suppose the characteristic of F is $p > 0$. Now

$$\mathbb{Z}/\text{Ker } f \simeq \mathcal{I}(f).$$

Since the characteristic of F is not zero, $\mathcal{I}(f) \neq \{0\}$. Therefore, $\mathcal{I}(f)$ is a nontrivial subring with 1 of the field F. Consequently, $\mathcal{I}(f)$ is an integral domain and so $\mathbb{Z}/\text{Ker } f$ is an integral domain. This implies Ker f is a prime ideal of $\mathbb{Z}$ and $\mathbb{Z} \neq$ Ker f. There exists a prime q such that Ker $f = q\mathbb{Z}$. Now $q1 = 0$ implies that $p|q$ and so $q = p$. Hence, $\mathbb{Z}/\text{Ker } f \simeq \mathbb{Z}_p$.

■

Let L be a subfield of $\mathbb{Q}$. Since $L\backslash\{0\}$ is a subgroup of $\mathbb{Q}\backslash\{0\}$ under multiplication, $1 \in L$. Hence, $\mathbb{Z} \subseteq L$ and so $\mathbb{Q} \subseteq L$. Thus, $\mathbb{Q}$ has no proper subfield. Similarly, $\mathbb{Z}_p$ has no proper subfield, where p is a prime.

Thus, the subfield K of the field F in Theorem 16.1.2 is the prime subfield of F.

The following theorem can be easily verified. We leave its proof as an exercise.

Theorem 16.1.3 *Let F be a field and K be a subfield of F. The following conditions are equivalent.*

(i) K is the prime subfield of F.

(ii) K is the intersection of all subfields of F. ■

Let F be a field and K a subfield of F. The field F is called an **extension** of the field K. We express this by F/K and call F/K a **field extension** or an **extension field**.

Definition 16.1.4 *Let F/K be a field extension and C be a subset of F. Define $K(C)$ to be the intersection of all subfields of F which contain $K \cup C$. Then the subfield $K(C)$ of F is called the subfield of F **generated** by C over K. C is called a set of **generators** for $K(C)/K$.*

Let $K[C]$ be the smallest subring of F containing $K \cup C$. Since any subfield of F which contains $K \cup C$ must contain $K[C]$, we have that $K(C)$ equals the intersection of all subfields which contain $K[C]$. Now $K[C]$ is an integral domain since it is a subring (with identity) of a field. Thus, by Theorem 9.1.6,

$$K(C) = \{ab^{-1} \mid a, b \in K[C],\ b \neq 0\}.$$

That is, $K(C)$ is the set of all rational expressions of the elements of $K[C]$. Hence, $K(C)$ is a quotient field of $K[C]$.

Let F/K be a field extension and $c_1, c_2, \ldots, c_n \in F$. Considering Definition 16.1.4, it follows that $K(c_1, c_2, \ldots, c_n) = K(c_1, c_2, \ldots, c_{n-1})(c_n)$. Recall that $K(c_1) = \{ab^{-1} \mid a,\ b \in K[c_1],\ b \neq 0\}$.

Definition 16.1.5 *Let F/K be a field extension. An element $a \in F$ is said to be **algebraic** over K if there exist $k_0, k_1, \ldots, k_n \in K$, not all zero, such that $k_0 + k_1 a + \cdots + k_n a^n = 0$; otherwise a is called **transcendental** over K.*

Let F/K be a field extension and let $a \in F$. Then a is algebraic over K if and only if a is a root of a nonzero polynomial with coefficients from K.

Example 16.1.6 *The element $\sqrt{2}$ in $\mathbb{R}$ is algebraic over $\mathbb{Q}$ since $\sqrt{2}$ is a root of $x^2 - 2 \in \mathbb{Q}[x]$. The element $i \in \mathbb{C}$ is algebraic over $\mathbb{R}$ and $\mathbb{Q}$ since i is a root of $x^2 + 1 \in \mathbb{Q}[x]$.*

Example 16.1.7 *It can be shown that π, $e \in \mathbb{R}$ are transcendental over $\mathbb{Q}$. In the quotient field $F(x)$ of the polynomial ring $F[x]$, F a field, x is transcendental over F since $\sum_{i=0}^{n} a_i x^i = 0$ if and only if $a_i = 0$ for $i = 0, 1, \ldots, n$.*

Theorem 16.1.8 *Let F/K be a field extension and $c \in F$. Then c is algebraic over K if and only if c is a root of some unique irreducible monic polynomial $p(x)$ over K.*

Proof. Suppose c is algebraic over K. There exists a nonzero polynomial $f(x) \in K[x]$ such that c is a root of $f(x)$ and $f(x) \notin K$. By Theorem 13.1.15, there exist irreducible polynomials $f_1(x)$, $f_2(x), \ldots, f_m(x) \in K[x]$ such that $f(x) = f_1(x) f_2(x) \cdots f_m(x)$. Thus,

$$0 = f(c) = f_1(c) f_2(c) \cdots f_m(c).$$

Since F has no zero divisors, we must have $f_i(c) = 0$ for some i. Thus, there exists an irreducible polynomial $h(x) = b_0 + b_1x + \cdots + b_mx^m$, $b_m \neq 0$, such that $h(c) = 0$. Let $p(x) = b_m^{-1}h(x)$. Then $p(x)$ is an irreducible monic polynomial in $K[x]$ with c as a root.

Let $g(x)$ be any polynomial in $K[x]$, which has c as a root. Let $p(x)$ be a monic polynomial of smallest degree in $K[x]$, which has c as a root. There exist $q(x)$, $r(x) \in K[x]$ such that $g(x) = q(x)p(x) + r(x)$, where either $r(x) = 0$ or $\deg r(x) < \deg p(x)$. Now

$$0 = g(c) = q(c)p(c) + r(c) = q(c) \cdot 0 + r(c).$$

Thus, $r(c) = 0$, whence $r(x) = 0$ else we contradict the minimality of the degree of $p(x)$. This implies that $p(x)|g(x)$ in $K[x]$. Let $s(x)$ be any irreducible polynomial in $K[x]$, which has c as a root (one such polynomial is $f_i(x)$ for some i, $1 \leq i \leq m$). Then $p(x)|s(x)$. Now $p(x)$ is not a constant polynomial in $K[x]$ since it has c as a root. Thus, since $s(x)$ is irreducible in $K[x]$, $p(x)$ must be irreducible in $K[x]$. Also, $p(x) = ks(x)$ for some $k \in K$. If we choose $s(x)$ monic, then $k = 1$ and so we have the desired uniqueness property of $p(x)$. The converse is immediate.

■

The proof of Theorem 16.1.8 yields the next result.

Corollary 16.1.9 *Let F/K be a field extension and $c \in F$ be such that c is algebraic over K. Then the unique monic irreducible polynomial $p(x)$ over K having c as a root satisfies the following properties:*

(i) There is no polynomial $g(x) \in K[x]$ having smaller degree than $p(x)$ and which has c as a root.

(ii) If c is a root of some $g(x) \in K[x]$, then $p(x)|g(x)$ in $K[x]$. ■

We call the polynomial $p(x)$, in Corollary 16.1.9, the **minimal polynomial** of c over K. The degree of $p(x)$ is called the **degree** of c over K

Example 16.1.10 *By Examples* 16.1.6, 12.3.6, *and* 12.3.7, *we have that $x^2 - 2$ is the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$ and $x^2 + 1$ is the minimal polynomial of i over $\mathbb{R}$.*

Theorem 16.1.11 *Let F/K be a field extension and $c \in F$.*

(i) If c is transcendental over K, then $K(c) \simeq K(x)$, where $K(x)$ is the quotient field of the polynomial ring $K[x]$.

(ii) If c is algebraic over K, then $K[c] \simeq K[x]/\langle p(x)\rangle$, where $p(x)$ is the minimal polynomial of c over K.

Proof. Define the mapping $\alpha : K[x] \longrightarrow K[c]$ by for all $f(x) \in K[x]$,

$$\alpha(f(x)) = f(c).$$

Then by Theorem 11.1.14, α is a homomorphism of $K[x]$ onto $K[c]$. Thus,

$$K[x]/\text{Ker } \alpha \simeq K[c].$$

(i) Now $f(x) \in \text{Ker } \alpha$ if and only if $f(c) = 0$, i.e., if and only if c is a root of $f(x)$. Hence, $\text{Ker } \alpha = \{0\}$ if and only if c is transcendental over K. Thus, c is transcendental over K implies α is an isomorphism of $K[x]$ onto $K[c]$ and so by Exercise 5 (page 168), α can be extended to an isomorphism of $K(x)$ onto $K(c)$. Consequently, if c is transcendental over K, then $K(x) \simeq K(c)$.

(ii) Suppose c is algebraic over K. Since $K[x]$ is a principal ideal domain, there exists $g(x) \in K[x]$ such that $\text{Ker } \alpha = \langle g(x)\rangle$. Now $\alpha(g(x)) = g(c) = 0$. Hence, c is a root of $g(x)$. Thus, $p(x)|g(x)$ and so there exists $q(x) \in K[x]$ such that $g(x) = q(x)p(x)$. This implies that $g(x) \in \langle p(x)\rangle$ and so

$$\text{Ker } \alpha = \langle g(x)\rangle \subseteq \langle p(x)\rangle.$$

Since $p(c) = 0$, $p(x) \in \text{Ker } \alpha$. Therefore, $\langle p(x)\rangle \subseteq \text{Ker } \alpha$. Consequently, $\text{Ker } \alpha = \langle p(x)\rangle$. ■

Corollary 16.1.12 *Let F/K be a field extension and $c \in F$. Then*

(i) $K[c] \subset K(c)$ if and only if c is transcendental over K,

(ii) $K[c] = K(c)$ if and only if c is algebraic over K.

Proof. Since $K[c] \subseteq K(c)$ always holds, (i) and (ii) are equivalent statements. Hence, we show that (ii) holds. Suppose c is algebraic over K. Then by Theorem 16.1.11,

$$K[c] \simeq K[x]/\langle p(x)\rangle$$

and since $p(x)$ is irreducible, $K[x]/\langle p(x)\rangle$ is a field. Thus, $K[c] = K(c)$. Conversely, suppose $K[c] = K(c)$. If $c = 0$, then c is the root of the polynomial $x \in K[x]$. Suppose that $c \neq 0$. Then $c^{-1} \in K(c)$ and so $c^{-1} = k_0 + k_1 c + \cdots + k_n c^n$ for some $k_i \in K$. This implies that $0 = -1 + k_0 c + k_1 c^2 + \cdots + k_n c^{n+1}$ and so c is algebraic over K.
■

Let F/K be a field extension. Under the field operations of F, F can be considered as a vector space over K. The elements of F are thought of as "vectors" while those of K are thought of as "scalars." Recall that $(F, +)$ is a commutative group and that for all $k_1, k_2 \in K$ and $a_1, a_2 \in F$, $k_1(a_1+a_2) = k_1a_1+k_1a_2$, $(k_1+k_2)a_1 = k_1a_1+k_2a_1$ hold from the distributive laws and that $(k_1k_2)a_1 = k_1(k_2a_1)$ holds from the associative law of multiplication.

Definition 16.1.13 *Let F/K be a field extension. The dimension of the vector space F over K is called the* ***degree*** *or* ***dimension*** *of F/K and is denoted by $[F : K]$. If the dimension of F/K is finite, then F/K is called a* ***finite extension****.*

Theorem 16.1.14 *Let F/K be a field extension and $c \in F$ be algebraic over K. Let $p(x)$ be the minimal polynomial of c over K. If $\deg p(x) = n$, then $\{1, c, c^2, \ldots, c^{n-1}\}$ is a basis of $K(c)/K$.*

Proof. By Corollary 16.1.12, $K[c] = K(c)$. Let $g(c) \in K[c]$ and $g(x)$ be the corresponding element in $K[x]$. There exist $q(x)$, $r(x) \in K[x]$ such that $g(x) = q(x)p(x) + r(x)$, where either $r(x) = 0$ or $\deg r(x) < \deg p(x)$. Thus, $g(c) = q(c)p(c) + r(c) = r(c)$. Hence, $\{1, c, c^2, \ldots, c^{n-1}\}$ spans $K(c)/K$. Suppose $0 = \sum_{i=0}^{n-1} k_i c^i$, $k_i \in K$. If the k_i's are not all zero, then c is a root of a polynomial of degree $\leq n-1 < n$, a contradiction. Thus, $k_i = 0$ for $i = 0, 1, \ldots, n-1$ and so $\{1, c, c^2, \ldots, c^{n-1}\}$ is linearly independent over K. Hence, $\{1, c, c^2, \ldots, c^{n-1}\}$ is a basis of $K(c)/K$. ■

Corollary 16.1.15 *Let F/K be a field extension. If $c \in F$ is algebraic and of degree n over K, then $[K(c) : K] = n$.* ■

Example 16.1.16 *The field extension $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$ is of degree 2 and $\{1, \sqrt{2}\}$ is a basis of $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$ since $p(x) = x^2 - 2$ is the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$ by Example 16.1.10. Thus, $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$.*

The student may recall from another mathematics course that $a + b\sqrt{2} = c + d\sqrt{2}$ if and only if $a = c$ and $b = d$, where $a, b, c, d \in \mathbb{Q}$. This becomes clear now since 1 and $\sqrt{2}$ are linearly independent over $\mathbb{Q}$ by Theorem 16.1.14.

Example 16.1.17 *By Theorem 16.1.14, the field extension $\mathbb{R}(i)/\mathbb{R}$ is of degree 2 and $\{1, i\}$ is a basis of $\mathbb{R}(i)$ over $\mathbb{R}$ since $p(x) = x^2 + 1$ is the minimal polynomial of i over $\mathbb{R}$. Thus, $\mathbb{R}(i) = \{a + bi \mid a, b \in \mathbb{R}\}$. Hence, we see that $\mathbb{R}(i)$ is $\mathbb{C}$, the field of complex numbers.*

Theorem 16.1.18 *Let F/K be a finite field extension. Then every element of F is algebraic over K.*

Proof. Let n be the dimension of F/K. Let $c \in F$ be such that $c \neq 0$, $c \neq 1$. (Clearly 0 and 1 are algebraic over K.) If the set $\{1, c, c^2, \ldots, c^n\}$ does not contain $n+1$ distinct elements, then $c^{j-i} = 1$ for some i, j $(0 \leq i < j \leq n)$ and so c is a root of $x^{j-i} - 1$. Suppose $1, c, c^2, \ldots, c^n$ are distinct. Then they must be linearly dependent since they are more in number than the dimension of the vector space F over K. Hence, there exist $k_0, k_1, \ldots, k_n \in K$ not all zero such that $0 = \sum_{i=0}^{n} k_i c^i$. Thus, c is a root of the polynomial $\sum_{i=0}^{n} k_i x^i$ over K.
■

The converse of Theorem 16.1.18 is not true, that is, it is not necessarily the case that if every element of F is algebraic over K, then F/K is a finite field extension. It can be shown that the set of all elements A of $\mathbb{C}$, which are algebraic over $\mathbb{Q}$ is a field such that $[A : \mathbb{Q}]$ is infinite (Theorem 16.1.22 and Example 16.1.25). A is called the **field of algebraic** numbers.

Theorem 16.1.19 *Let $K(c)/K$ be a field extension. Then $K(c)/K$ is finite if and only if c is algebraic over K.*

Proof. If $K(c)/K$ is finite, then c is algebraic over K by Theorem 16.1.18. If c is algebraic over K, then $K(c)/K$ is finite by Corollary 16.1.15.
■

Let F/K be a field extension. A subfield L of F is called an **intermediate field** of F/K if $K \subseteq L \subseteq F$. Since $a - b \in L$ for all $a, b \in L$ and $ka \in L$ for all $k \in K$ and $a \in L$, it follows that L is a subspace of F over K. An intermediate field L of F/K is called **proper** if $L \neq F$.

Theorem 16.1.20 *Let F/K be a field extension and L be an intermediate field of F/K. Then*

$$[F : K] = [F : L][L : K].$$

Moreover, F/K is a finite extension if and only if F/L and L/K are finite extensions.

Proof. Let V be a basis of F/L and U be a basis of L/K. We show that

$$W = \{uv \mid u \in U, v \in V\}$$

is a basis of F/K. Let $c \in F$. Since V is a basis of F/L, there exist $v_1, v_2, \ldots, v_n \in V$ and $c_1, c_2, \ldots, c_n \in L$ such that

$$c = \sum_{j=1}^{n} c_j v_j. \tag{16.1}$$

Since U is a basis of L/K, there exist $u_1, u_2, \ldots, u_m \in U$ and $k_{1j}, k_{2j}, \ldots, k_{mj} \in K$ such that

$$c_j = \sum_{i=1}^{m} k_{ij} u_i, \qquad j = 1, 2, \ldots, n. \tag{16.2}$$

Substituting Eq. (16.2) into Eq. (16.1), we obtain

$$c = \sum_{j=1}^{n} \sum_{i=1}^{m} k_{ij} u_i v_j.$$

Thus, W spans F over K. Suppose

$$0 = \sum_{j=1}^{n} \sum_{i=1}^{m} k_{ij} u_i v_j,$$

where $u_i \in U$, $v_j \in V$, and $k_{ij} \in K$ for all $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$. Then

$$0 = \sum_{j=1}^{n} (\sum_{i=1}^{m} k_{ij} u_i) v_j$$

and since V is linearly independent over L,

$$0 = \sum_{i=1}^{m} k_{ij} u_i, \ j = 1, 2, \ldots, n.$$

Thus, $k_{ij} = 0$ for $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$ since U is linearly independent over K. Hence, W is linearly independent over K, whence W is a basis of F over K. Let $u, u' \in U$ and $v, v' \in V$. If $v \neq v'$, then $uv \neq u'v'$ since v and v' are linearly independent over L. If $v = v'$, then $uv = u'v'$ if and only if $u = u'$. Consequently, for all $u, u' \in U$ and for all $v, v' \in V$ if either $u \neq u'$ or $v \neq v'$, then $uv \neq u'v'$. Hence, $[F : K] = |U \times V| = |U|\,|V| = [F : L][L : K]$. Now if either U or V is infinite, then W is infinite. If U and V are finite sets, then W is a finite set. Hence, F/K is a finite extension if and only if F/L and L/K are finite extensions. ■

Example 16.1.21 *Consider the field extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}$. By Example* 16.1.10*, $x^2 - 2$ is the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$. Also, $x^2 - 3$ is the minimal polynomial of $\sqrt{3}$ over $\mathbb{Q}(\sqrt{2})$. (That $x^2 - 3$ is irreducible over $\mathbb{Q}(\sqrt{2})$ follows by an argument that is similar to the one used in Worked-Out Exercise* 1*, page* 235*.) Thus, $\{1, \sqrt{2}\}$ is a basis of $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$ and $\{1, \sqrt{3}\}$ is a basis of $\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}(\sqrt{2})$. By Theorem* 16.1.20*, $\{1, \sqrt{2}, \sqrt{3}, \sqrt{6}\}$ is a basis of $\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}$. $[\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}] = 4$, $[\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}(\sqrt{2})] = 2$, and $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$.*

Theorem 16.1.22 *Let F/K be a field extension. If L is the set of all elements in F, which are algebraic over K, then L is an intermediate field of F/K.*

Proof. Any $k \in K$ is a root of the polynomial $x - k$ over K. Thus, $L \supseteq K$. Let a and b be elements of L, where a is of degree m over K and b is of degree n over K. Then $K(a)/K$ is of degree m and $K(a,b)/K(a)$ is of degree at most n. Hence, by Theorem 16.1.20, $K(a,b)/K$ is a finite extension. By Theorem 16.1.18, every element of $K(a,b)$ is algebraic over K. Since $a-b$ and ab^{-1} (for $b \neq 0$) are elements of $K(a,b)$, $a-b$ and ab^{-1}(for $b \neq 0$) are algebraic over K. Thus, $a - b$ and ab^{-1} (for $b \neq 0$) $\in L$ and so L is a field. ■

Definition 16.1.23 *A field extension F/K is called* ***algebraic*** *if every element of F is algebraic over K; otherwise F/K is called* ***transcendental.***

Theorem 16.1.24 *Let L be an intermediate field of the field extension F/K. Then F/K is an algebraic extension if and only if F/L and L/K are algebraic extensions.*

Proof. Suppose that F/K is algebraic. Let $a \in F$. Then a is a root of a nonzero polynomial $p(x) \in K[x]$. Since $K \subseteq L$, $p(x) \in L[x]$. Thus, a is algebraic over L and so F/L is algebraic. Every element of L is an element of F. Hence, L/K is algebraic. Conversely, suppose F/L and L/K are algebraic extensions. Let $c \in F$. Then c is a root of some nonzero polynomial $c_0 + c_1x + \cdots + c_nx^n \in L[x]$. Thus, c is algebraic over $K(c_0, c_1, \ldots, c_n)$ whence $K(c_0, c_1, \ldots, c_n)(c)/K(c_0, c_1, \ldots, c_n)$ is a finite extension. Since $c_0, c_1, \ldots, c_n$ are algebraic over K, repeated application of Theorem 16.1.20 yields that $K(c_0, c_1, \ldots, c_n)(c)/K$ is a finite extension. Therefore, c is algebraic over K by Theorem 16.1.18. Hence, F/K is an algebraic extension. ■

Example 16.1.25 *Let $F = \mathbb{Q}(\{\sqrt{p} \mid p \in \mathbb{Z}, p \text{ is a prime}\}) \subseteq \mathbb{R}$. We show that $F/\mathbb{Q}$ is algebraic and $[F : \mathbb{Q}] = \infty$. Now for any prime p, $\sqrt{p} \notin \mathbb{Q}$. Let $p_1, \ldots, p_n$ be any distinct primes. Suppose $p \neq p_i$, $i = 1, 2, \ldots, n$, and p is a prime. Assume that $\sqrt{p} \notin \mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_n})$, the induction hypothesis. (The case $n = 0$ is $\sqrt{p} \notin \mathbb{Q}$ and this case is described above.) We show that if $p_1, \ldots, p_{n+1}$ are distinct primes and $p \neq p_i$, $i = 1, 2, \ldots, n+1$, then $\sqrt{p} \notin \mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_{n+1}})$. Suppose $\sqrt{p} \in \mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_{n+1}})$. Then there exist $a, b \in \mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_n})$ such that $\sqrt{p} = a + b\sqrt{p_{n+1}}$. If $a = 0$, then $p = b^2p_{n+1}$, a contradiction since p and p_{n+1} are distinct primes. If $b = 0$, then $\sqrt{p} = a \in \mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_n})$, a contradiction to our induction hypothesis. Suppose $a \neq 0$ and $b \neq 0$. Then $p = a^2 + p_{n+1}b^2 + 2ab\sqrt{p_{n+1}}$. Hence, $\sqrt{p_{n+1}} = (p - a^2 - p_{n+1}b^2)/2ab \in \mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_n})$ and so $\sqrt{p} \in \mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_n})$, a contradiction of the hypothesis. Hence, $\sqrt{p} \notin \mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_{n+1}})$. Thus, by the induction hypothesis, we find that for any positive integer k, if $p_1, \ldots, p_k, p$ are distinct primes, then $\sqrt{p} \notin \mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_k})$. Hence,*

$$\mathbb{Q} \subset \mathbb{Q}(\sqrt{2}) \subset \mathbb{Q}(\sqrt{2}, \sqrt{3}) \subset \cdots$$

is an infinite strictly ascending chain of intermediate fields of $F/\mathbb{Q}$. Hence, $F/\mathbb{Q}$ must be of infinite dimension. Let $a \in F$. Then there exist primes $p_1, \ldots, p_n$ such that $a \in \mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_n})$. Since $\mathbb{Q}(\sqrt{p_1}, \ldots, \sqrt{p_n})/\mathbb{Q}$ is a finite field extension, a is algebraic over $\mathbb{Q}$ by Theorem 16.1.18. *Hence, $F/\mathbb{Q}$ is algebraic. Note that from this example, it follows that $[\mathbb{R} : \mathbb{Q}] = \infty$.*

The above example provides us with a field extension $F/\mathbb{Q}$ which shows that the converse of Theorem 16.1.18 is not true. Since the field of algebraic numbers A contains F, we have $[A : \mathbb{Q}] = \infty$.

Definition 16.1.26 *Let F/K and L/K be field extensions and $\sigma : F \to L$ be a homomorphism. Then σ is called a K-**homomorphism** if $\sigma(a) = a$ for all $a \in K$.*

Let F/K and L/K be field extensions and $\sigma : F \to L$ be a K-homomorphism. Since σ is a nonzero homomorphism, Ker $\sigma \neq F$. Therefore, Ker $\sigma = \{0\}$ since the only ideals of F are F and $\{0\}$. This implies that σ is one-one. Hence, σ is an isomorphism of F onto $\sigma(F)$. We simply call σ a K-**isomorphism** of F into L. If $L = F = \sigma(F)$ and σ is a K-isomorphism of F into L, then we call σ a K-**automorphism**.

Theorem 16.1.27 *Let F/K be an algebraic extension and $\sigma : F \to F$ be a K-homomorphism. Then σ is an automorphism.*

Proof. As above σ is one-one. To show σ is an automorphism, it only remains to be shown that $\sigma(F) = F$, i.e., σ is onto F.Let $a \in F$. Let $f(x) = a_0 + a_1x + \cdots + a_kx^k \in K[x]$ be the minimal polynomial of a over K. Let b be any root of $f(x)$ in F. Then $f(\sigma(b)) = a_0 + a_1\sigma(b) + \cdots + a_k\sigma(b)^k = \sigma(a_0 + a_1b + \cdots + a_kb^k) = 0$. Hence, $\sigma(b)$ is a root of $f(x)$. Let F' be the subfield of F generated by all roots of $f(x)$ over K that lie in F. Then F'/K is a finite extension. Since σ maps a root of $f(x)$ to a root of $f(x)$, σ maps F' into F'. Since $[F' : K] = [\sigma(F') : K]$, it now follows that $[F' : \sigma(F')] = 1$ by Theorem 16.1.20 and so $F' = \sigma(F')$. Hence, $a \in F' = \sigma(F') \subseteq \sigma(F)$. Thus, σ is onto F. ■

Worked-Out Exercises

♢ **Exercise 1**: Show that the polynomial $x^2 - 7$ is irreducible in $\mathbb{Q}(\sqrt{3})[x]$.

Solution: Suppose $x^2 - 7 = (x - (a + b\sqrt{3}))(x - (c + d\sqrt{3}))$, where $a, b, c, d \in \mathbb{Q}$. Then $x^2 - 7 = x^2 - ((a + c) + (b + d)\sqrt{3})x + (ac + 3bd + ad\sqrt{3} + bc\sqrt{3})$. This implies that

$$\begin{aligned} (a + c) + (b + d)\sqrt{3} &= 0 \\ ac + 3bd + ad\sqrt{3} + bc\sqrt{3} &= -7. \end{aligned}$$

Since $\{1, \sqrt{3}\}$ is linearly independent over $\mathbb{Q}$, $a + c = 0$ and $b + d = 0$. Hence,

$$-a^2 - 3b^2 + (-2ab)\sqrt{3} = -7.$$

Thus, $-a^2 - 3b^2 = -7$ and $-2ab = 0$. Hence, $ab = 0$. Suppose $a = 0$. Then $3b^2 = 7$. Now $b = \frac{m}{n}$ for some integers m and n with $\gcd(m, n) = 1$. Therefore, $3m^2 = 7n^2$, which contradicts the fundamental theorem of arithmetic. Suppose $b = 0$. Then $a^2 = 7$, which again leads to a contradiction of the fundamental theorem of arithmetic. Thus, $x^2 - 7$ is irreducible in $\mathbb{Q}(\sqrt{3})[x]$.

♢ **Exercise 2**: Find $[\mathbb{Q}(\sqrt{3}, \sqrt{7}) : \mathbb{Q}(\sqrt{3})]$ and $[\mathbb{Q}(\sqrt{3}) : \mathbb{Q}]$. Also, find a basis for $\mathbb{Q}(\sqrt{3}, \sqrt{7})/\mathbb{Q}(\sqrt{3})$ and a basis for $\mathbb{Q}(\sqrt{3}, \sqrt{7})/\mathbb{Q}$.

Solution: By Worked-Out Exercise 1 (page 235), $x^2 - 7$ is irreducible over $\mathbb{Q}(\sqrt{3})$. Thus,

$$[\mathbb{Q}(\sqrt{3}, \sqrt{7}) : \mathbb{Q}(\sqrt{3})] = \deg(x^2 - 7) = 2.$$

By Theorem 16.1.14, $\{1, \sqrt{7}\}$ is a basis for $\mathbb{Q}(\sqrt{3}, \sqrt{7})/\mathbb{Q}(\sqrt{3})$. Since $x^2 - 3$ is irreducible over $\mathbb{Q}$, $[\mathbb{Q}(\sqrt{3}) : \mathbb{Q}] = 2$ and $\{1, \sqrt{3}\}$ is a basis for $\mathbb{Q}(\sqrt{3})/\mathbb{Q}$. Thus,

$$[\mathbb{Q}(\sqrt{3}, \sqrt{7}) : \mathbb{Q}] = [\mathbb{Q}(\sqrt{3}, \sqrt{7}) : \mathbb{Q}(\sqrt{3})][\mathbb{Q}(\sqrt{3}) : \mathbb{Q}] = 2 \cdot 2 = 4.$$

By Theorem 16.1.20, $\{1, \sqrt{3}, \sqrt{7}, \sqrt{21}\}$ is a basis of $\mathbb{Q}(\sqrt{3}, \sqrt{7})/\mathbb{Q}$.

♢ **Exercise 3**: Find an element $u \in \mathbb{R}$ such that $\mathbb{Q}(\sqrt{2}, \sqrt[3]{7}) = \mathbb{Q}(u)$.

Solution: We claim that $u = \sqrt{2}\sqrt[3]{7}$. Since $u = \sqrt{2}\sqrt[3]{7} \in \mathbb{Q}(\sqrt{2}, \sqrt[3]{7})$, $\mathbb{Q}(u) \subseteq \mathbb{Q}(\sqrt{2}, \sqrt[3]{7})$. Now $\sqrt{2}\sqrt[3]{7} \in \mathbb{Q}(u)$ implies that $14\sqrt{2} = (\sqrt{2}\sqrt[3]{7})^3 \in \mathbb{Q}(u)$. Hence, $\sqrt{2} \in \mathbb{Q}(u)$. Since $\sqrt{2}, \sqrt{2}\sqrt[3]{7} \in \mathbb{Q}(u)$, $\sqrt[3]{7} \in \mathbb{Q}(u)$. Therefore, $\mathbb{Q}(\sqrt{2}, \sqrt[3]{7}) \subseteq \mathbb{Q}(u)$. Thus, $\mathbb{Q}(\sqrt{2}, \sqrt[3]{7}) = \mathbb{Q}(u)$.

♢ **Exercise 4**:
(a) Let F be a field and a, b be members of a field containing F. Suppose that a and b are algebraic of degree m and n over F, respectively. Suppose m and n are relatively prime. Show that $[F(a, b) : F] = mn$.

(b) Show that the result in (i) need not be true if m and n are not relatively prime.

Solution:
(a) Let $f(x) \in F[x]$ be the minimal polynomial of a of degree m. Now $f(x) \in F[x] \subseteq F(b)[x]$. Thus, a satisfies a polynomial of degree m over $F(b)$. Hence, $[F(b)(a) : F(b)] \leq m$. Since $F(b)(a) = F(a, b)$, $[F(a, b) : F(b)] \leq m$. Now $[F(a, b) : F] = [F(a, b) : F(b)][F(b) : F] \leq mn$. Also,

$$[F(a, b) : F] = [F(a, b) : F(b)][F(b) : F] = [F(a, b) : F(b)]n.$$

Thus, $n|[F(a, b) : F]$. Similarly, $m|[F(a, b) : F]$. Since m and n are relatively prime, $mn|[F(a, b) : F]$. Therefore, $[F(a, b) : F] \geq mn$. Consequently, $[F(a, b) : F] = mn$.

(b) Let $F = \mathbb{Q}$, $a = 2^{\frac{1}{6}}$, and $b = 2^{\frac{1}{4}}$. Then a is algebraic over F of degree 6 and b is algebraic over F of degree 4. We claim that $F(a, b) = F(2^{\frac{1}{12}})$. Now $b = (2^{\frac{1}{12}})^3 \in F(2^{\frac{1}{12}})$ and $a = (2^{\frac{1}{12}})^2 \in F(2^{\frac{1}{12}})$. Thus, $F(a, b) \subseteq F(2^{\frac{1}{12}})$. Now $2^{\frac{1}{12}} = 2^{\frac{1}{4} - \frac{1}{6}} = 2^{\frac{1}{4}}(2^{\frac{1}{6}})^{-1} \in F(a, b)$. Hence, $F(a, b) = F(2^{\frac{1}{12}})$. Since $x^{12} - 2$ is the minimal polynomial of $2^{\frac{1}{12}}$, $[F(2^{\frac{1}{12}}) : F] = 12 \neq 24 = 4 \cdot 6$.

♢ **Exercise 5**: Consider the unique factorization domain $F[t]$, where F is a field and t is transcendental over F. Show that the polynomial $x^2 + tx + t \in F(t)[x]$ is irreducible over $F(t)$. Also, show that $x^2 + tx + t \in F(x)[t]$ is irreducible over $F(x)$.

Solution: Now $t \not| 1, t|t$, but $t^2 \not| t$. Note t is prime in $F[t]$. Thus, $x^2 + tx + t \in F(t)[x]$ is irreducible over $F(t)$ by Eisenstein's criterion. If we consider $x^2 + tx + t$ as a polynomial in t over $F(x)$, then $x^2 + tx + t = (x + 1)t + x^2$. It follows that Eisenstein's criterion does not apply. However, since $(x + 1)t + x^2$ is of degree 1 in t, it is irreducible over $F(x)$.

Exercise 6: Let $K[u,v]$ denote the polynomial ring in two algebraic independent indeterminates u, v over the field K. Let F denote the field of quotients $K(u,v)$ of $K[u,v]$. Prove that the polynomial x^2+vx+u is irreducible over F.

Solution: Suppose x^2+vx+u is reducible over F. Then

$$x^2+vx+u = \left(x+\frac{p(u,v)}{q(u,v)}\right)\left(x+\frac{f(u,v)}{g(u,v)}\right),$$

where $p(u,v)$, $q(u,v)$, $f(u,v)$, $g(u,v) \in K[u,v]$. We may assume that $p(u,v)$ and $q(u,v)$ are relatively prime in $K[u,v]$ and also $f(u,v)$ and $g(u,v)$ are relatively prime in $K[u,v]$. Now

$$uq(u,v)g(u,v) = p(u,v)f(u,v). \tag{16.3}$$

Hence, $g(u,v)$ divides $p(u,v)$, $p(u,v)$ divides $ug(u,v)$, $q(u,v)$ divides $f(u,v)$, and $f(u,v)$ divides $uq(u,v)$. Also,

$$v = \frac{p(u,v)}{q(u,v)} + \frac{f(u,v)}{g(u,v)}.$$

Consequently,

$$vq(u,v)g(u,v) = p(u,v)g(u,v) + q(u,v)f(u,v). \tag{16.4}$$

Therefore, $g(u,v)$ divides $q(u,v)$ and $q(u,v)$ divides $g(u,v)$. Thus,

$$g(u,v) = kq(u,v)$$

for some $k \in K$. Hence, $g(u,v)$ and $p(u,v)$ are relatively prime. Similarly, $q(u,v)$ and $f(u,v)$ are relatively prime. Thus, $p(u,v)$ divides u and $f(u,v)$ divides u by Eq. (16.3). Hence,

$$\text{either } p(u,v) = k_1 u \text{ or } p(u,v) = k_1, \tag{16.5}$$

$$\text{either } f(u,v) = k_2 u \text{ or } f(u,v) = k_2 \tag{16.6}$$

for some $k_1, k_2 \in K$. Suppose that $p(u,v) = k_1 u$ and $f(u,v) = k_2 u$. Then substituting into Eq. (16.4) we obtain

$$vq(u,v)g(u,v) = k_1 ug(u,v) + k_2 uq(u,v).$$

Thus,

$$vq(u,v)g(u,v) = k_1 ukq(u,v) + k_2 uq(u,v).$$

Hence, $vg(u,v) = (k_1 k + k_2)u$. However, this contradicts the algebraic independence of u, v over K. Substituting the remaining possibilities in Eqs. (16.5) and (16.6) into Eq. (16.4), we also obtain a contradiction of the algebraic independence of u, v over K. Thus, x^2+vx+u is irreducible over F.

Exercise 7: Let $F = K(x,y)$, where K is a field and x, y are algebraically independent indeterminates over K. Show that $F \neq K(x)K(y)$, where

$$\begin{aligned} K(x)K(y) \quad = \quad & \{\textstyle\sum_i (p_i(x)/q_i(x))(u_i(y)/v_i(y)) \mid p_i(x), q_i(x) \in K[x], \\ & u_i(y), v_i(y) \in K[y], q_i(x) \neq 0, v_i(y) \neq 0\}. \end{aligned}$$

Solution: Now $\frac{1}{x+y} \notin K(x)K(y)$ else $\frac{1}{x+y} = (\sum_i (f_i(x)g_i(y))/h(x)k(y)$, after obtaining a common denominator. Thus,

$$h(x)k(y) = (x+y)(\sum_i (f_i(x)g_i(y)).$$

This implies that $x+y$ divides $h(x)k(y)$. Hence, $x+y$ divides $h(x)$ or $k(y)$ since $x+y$ is prime in the UFD $K[x,y]$, a contradiction of the algebraic independence of x, y over K.

Exercises

1. Show that $\mathbb{Q}(\sqrt{3}, -\sqrt{3}) = \mathbb{Q}(\sqrt{3})$.
2. Let F/K be a field extension. Show that $[F:K] = 1$ if and only if $F = K$.
3. Consider the field extension $\mathbb{R}/\mathbb{Q}$.
   (i) Show that π^2 is transcendental over $\mathbb{Q}$.
   (ii) Show that $\sqrt{\pi}$ is transcendental over $\mathbb{Q}$.
4. Consider the field extension $\mathbb{R}/\mathbb{Q}$. Show that $\pi - 3$ is transcendental over $\mathbb{Q}$.

5. Consider the field extension $\mathbb{R}/\mathbb{Q}$. Show that π is transcendental over $\mathbb{Q}(\sqrt{2})$.
6. Consider the field extension $\mathbb{R}/\mathbb{Q}$. Show that $\pi + \sqrt{2}$ is transcendental over $\mathbb{Q}$.
7. Let F/K be a field extension such that $[F : K] < \infty$. Let $p(x)$ be an irreducible polynomial in $K[x]$. Suppose $p(c) = 0$ for some $c \in F$. Prove that $\deg p(x)$ divides $[F : K]$.
8. Find $[\mathbb{Q}(\sqrt[3]{5}) : \mathbb{Q}]$.
9. Show that $\mathbb{Q}(\sqrt{3} - \sqrt{5}) = \mathbb{Q}(\sqrt{3}, \sqrt{5})$. Find $[\mathbb{Q}(\sqrt{3} - \sqrt{5}) : \mathbb{Q}]$.
10. Show that the polynomial $x^2 - 5$ is irreducible over $\mathbb{Q}(\sqrt{2})$.
11. Find the minimal polynomial of $\sqrt{2 + \sqrt{5}}$ over $\mathbb{Q}$.
12. Let $c = \sqrt[5]{3}$. Show that $\mathbb{Q}(c) = \mathbb{Q}(c^2)$.
13. Find $[\mathbb{Q}(\sqrt{2}, \sqrt{5}) : \mathbb{Q}(\sqrt{2})]$, $[\mathbb{Q}(\sqrt{2}, \sqrt{5}) : \mathbb{Q}]$, a basis for $\mathbb{Q}(\sqrt{2}, \sqrt{5})/\ \mathbb{Q}(\sqrt{2})$, and a basis for $\mathbb{Q}(\sqrt{2}, \sqrt{5})/\mathbb{Q}$.
14. Let F/K be a field extension and $c \in F$ be algebraic over K. Let $f(x) \in K[x]$. Show that $f(c)$ is algebraic over K.
15. Prove that if $[F : K] = p$, p a prime, then F/K has no proper intermediate fields.
16. Let L and M be intermediate fields of the field extension F/K. Suppose that $[L : K]$ is a prime. Prove that either $L \cap M = K$ or $L \subseteq M$.
17. Let F/K be a field extension, $f(x)$ be a nonzero polynomial in $K[x]$, and $c \in F$. If $f(x)$ is algebraic over K, prove that c is algebraic over K.
18. Let F/K be a field extension such that $[F : K] = p$, p a prime. Prove that if $c \in F$, $c \notin K$, then $F = K(c)$.
19. Let F/K be a field extension and a, $b \in F$ be algebraic over K. If a has degree m over K and $b \neq 0$ has degree n over K, prove that the elements $a + b$, ab, $a - b$, ab^{-1} have degree at most mn over K.
20. Prove that $\sqrt{2} + \sqrt{3}$, $\sqrt{2} - \sqrt{3}$ have degree 4 over $\mathbb{Q}$ and that $\sqrt{2}\sqrt{3}$, $\sqrt{2}/\sqrt{3}$ have degree 2 over $\mathbb{Q}$. Find the minimal polynomials of these elements over $\mathbb{Q}$.
21. Let F/K be a field extension and R be a ring such that $K \subseteq R \subseteq F$. Prove that if every element of R is algebraic over K, then R is a field.
22. Let F/K be a field extension and $u, v \in F$.
    (i) Prove that $K(u, u + v) = K(u, v)$.
    (ii) If u and $u + v$ are algebraic over K, prove that $[K(u, v) : K]$ is finite and v is algebraic over K.
23. Answer the following statements true or false. If the statement is true, prove it. If it is false, give a counterexample.
    (i) Let F/K be a field extension and L be an intermediate field of F/K. Let V be a basis of F/L such that $1 \in V$ and U be a basis of L/K such that $1 \in U$. Then $U \cup V$ is linearly independent over K.
    (ii) Let F/K be a field extension and L be an intermediate field of F/K. Let V be a basis of F/L and U be a basis of L/K. Then $U \cup V$ is a basis of F/K.
    (iii) Let F/K be a field extension and $c, d \in F$. If $K(c, d) = K(c)$, then $d = f(c)$ for some polynomial $f(x) \in K[x]$.

16.2 Splitting Fields

Here we give some results concerning the existence of field extensions which are generated by roots of polynomials. These results are basic to Galois theory.

Consider the polynomial ring $K[x]$ over the field K. Let $f(x) \in K[x]$. In the quotient ring $K[x]/\langle f(x)\rangle$, we let $\overline{g(x)}$ denote the coset $g(x) + \langle f(x)\rangle$. Thus, if $g(x) = \sum_{i=0}^{n} k_i x^i$, then by the definition of addition and multiplication of cosets, we have that $\overline{g(x)} = \sum_{i=0}^{n} \overline{k_i}\overline{x}^i$.

Theorem 16.2.1 (Kronecker) *Let K be a field. If $f(x)$ is a nonconstant polynomial in $K[x]$, then there exists a field extension F/K such that F contains a root of $f(x)$.*

Proof. Since $K[x]$ is a unique factorization domain, there exist irreducible polynomials $f_1(x), \ldots, f_n(x) \in K[x]$ such that $f(x) = f_1(x) \cdots f_n(x)$. Thus, a root of any $f_i(x)$, $i = 1, 2, \ldots, n$, is a root of $f(x)$. Hence, it suffices to prove the theorem for $f(x)$ irreducible in $K[x]$. The ideal $\langle f(x)\rangle$ is maximal in $K[x]$ and so $F = K[x]/\langle f(x)\rangle$ is a field. Let α be the natural homomorphism of $K[x]$ onto $K[x]/\langle f(x)\rangle$. Since $K \cap \langle f(x)\rangle = \{0\}$, α maps K

one-one into F. Thus, say, $K \subseteq F$, that is, we identify $k \in K$ with $\overline{k}$ in F. Hence, $\alpha(f(x)) = \overline{f(x)} = f(\overline{x})$, where $\overline{f(x)} = f(x) + \langle f(x) \rangle$ and $\overline{x} = x + \langle f(x) \rangle$. Now $\alpha(f(x)) = \overline{0}$ and so $f(\overline{x}) = \overline{0}$. Therefore, $\overline{x}$ is a root of $f(x)$.
■

The field extension F/K in Theorem 16.2.1 has some interesting properties. Consider the subring $K[\overline{x}]$ of F. Then $\alpha(\sum_{i=0}^{m} k_i x^i) = \sum_{i=0}^{m} k_i \overline{x}^i$ for all $\sum_{i=0}^{m} k_i x^i \in K[x]$ and so α maps $K[x]$ onto $K[\overline{x}]$. Since α also maps $K[x]$ onto F, we have $F = K[\overline{x}] = K(\overline{x})$. Thus, for $f(x)$ irreducible in $K[x]$, we have by Theorem 16.1.14 that $[F : K] = n$ and $\{1, \overline{x}, \ldots, \overline{x}^{n-1}\}$ is a basis of F/K, where $n = \deg f(x)$.

Example 16.2.2 *$x^2 + 1$ is irreducible in $\mathbb{R}[x]$. Now $\mathcal{C} = \mathbb{R}/\langle x^2 + 1 \rangle = \mathbb{R}[\overline{x}] = \{a + b\overline{x} \mid a, b \in \mathbb{R}\}$ is a field, where $\overline{x} = x + \langle x^2 + 1 \rangle$. Since $x^2 = -1$, we may call $\mathcal{C}$ the field of complex numbers. We may think of $\overline{x}$ as i.*

Example 16.2.3 *Consider the polynomial $x^4 - 3 \in \mathbb{Q}[x]$. By Eisenstein's criterion, $x^4 - 3$ is irreducible in $\mathbb{Q}[x]$. Set $\lambda = x + \langle x^4 - 3 \rangle$ in the field $\mathbb{Q}[x]/\langle x^4 - 3 \rangle$. Then*

$$\mathbb{Q}[x]/\langle x^4 - 3 \rangle = \mathbb{Q}(\lambda) = \{a + b\lambda + c\lambda^2 + d\lambda^3 \mid a, b, c, d \in \mathbb{Q}\}$$

and $\{1, \lambda, \lambda^2, \lambda^3\}$ is a basis of $\mathbb{Q}(\lambda)$ over $\mathbb{Q}$. Let us multiply two elements of $\mathbb{Q}(\lambda)$ and determine the form $a + b\lambda + c\lambda^2 + d\lambda^3$ for their product. Consider $(1 + \lambda + \lambda^3)$ and $(1 + \lambda^2)$. Then

$$(1 + \lambda + \lambda^3)(1 + \lambda^2) = 1 + \lambda + \lambda^2 + 2\lambda^3 + \lambda^5.$$

Now

$$1 + x + x^2 + 2x^3 + x^5 = x(x^4 - 3) + 1 + 4x + x^2 + 2x^3$$

using the division algorithm. Thus,

$$\begin{aligned} 1 + \lambda + \lambda^2 + 2\lambda^3 + \lambda^5 &= \lambda(\lambda^4 - 3) + 1 + 4\lambda + \lambda^2 + 2\lambda^3 \\ &= \lambda \cdot 0 + 1 + 4\lambda + \lambda^2 + 2\lambda^3. \end{aligned}$$

Hence,

$$(1 + \lambda + \lambda^3)(1 + \lambda^2) = 1 + 4\lambda + \lambda^2 + 2\lambda^3.$$

Let us find $(1 + \lambda + \lambda^3)^{-1}$. Since $x^4 - 3$ is irreducible over $\mathbb{Q}$, the gcd of $x^4 - 3$ and $x^3 + x + 1$ is 1. Therefore, there exist $s(x)$, $t(x) \in \mathbb{Q}[x]$ such that

$$1 = s(x)(x^4 - 3) + t(x)(1 + x + x^3).$$

Thus,

$$\begin{aligned} 1 &= s(\lambda)(\lambda^4 - 3) + t(\lambda)(1 + \lambda + \lambda^3) \\ 1 &= 0 + t(\lambda)(1 + \lambda + \lambda^3). \end{aligned}$$

Hence, $t(\lambda) = (1 + \lambda + \lambda^3)^{-1}$. We have not really calculated $t(\lambda)$, however. To do this calculation, we must know the exact form of $s(x)$ and $t(x)$. The method for finding $s(x)$ and $t(x)$ is described below. Now by repeated use of the division algorithm, we have

$$x^4 - 3 = x(x^3 + x + 1) + (-x^2 - x - 3)$$

$$\begin{aligned} x^3 + x + 1 &= (-x + 1)(-x^2 - x - 3) + (-x + 4) \\ -x^2 - x - 3 &= (x + 5)(-x + 4) + (-23) \\ -x + 4 &= (\tfrac{1}{23}x - \tfrac{4}{23})(-23) + 0. \end{aligned}$$

Thus, by back substitution, we obtain

$$\begin{aligned} -23 &= -x^2 - x - 3 - (x + 5)(-x + 4) \\ -23 &= -x^2 - x - 3 - (x + 5)[x^3 + x + 1 - (-x + 1)(-x^2 - x - 3)] \\ &= (-x^2 - 4x + 6)(-x^2 - x - 3) - (x + 5)(x^3 + x + 1) \\ &= (-x^2 - 4x + 6)[x^4 - 3 - x(x^3 + x + 1)] - (x + 5)(x^3 + x + 1) \\ &= (-x^2 - 4x + 6)(x^4 - 3) + (x^3 + 4x^2 - 7x - 5)(x^3 + x + 1). \end{aligned}$$

This implies that

$$1 = -\frac{1}{23}(-x^2 - 4x + 6)(x^4 - 3) + (-\frac{1}{23})(x^3 + 4x^2 - 7x - 5)(x^3 + x + 1).$$

Therefore,

$$t(x) = -\frac{1}{23}x^3 - \frac{4}{23}x^2 + \frac{7}{23}x + \frac{5}{23}.$$

Consequently,

$$(1+\lambda+\lambda^3)^{-1} = \frac{5}{23} + \frac{7}{23}\lambda - \frac{4}{23}\lambda^2 - \frac{1}{23}\lambda^3.$$

Since λ is a root of x^4-3 in $\mathbb{Q}(\lambda)$, we know by Corollary 11.1.10 *that $x-\lambda$ divides x^4-3 over $\mathbb{Q}(\lambda)$. In fact, $x^4-3=(x-\lambda)(x^3+\lambda x^2+\lambda^2 x+\lambda^3)$. We know there exists a field $\mathbb{Q}(\lambda)(\lambda_2)$, where λ_2 is a root of $x^3+\lambda x^2+\lambda^2 x+\lambda^3$ over $\mathbb{Q}(\lambda)$ by Theorem* 16.2.1. *Over the field $\mathbb{Q}(\lambda)(\lambda_2)$, $x^3+\lambda x^2+\lambda^2 x+\lambda^3$ factors into $(x-\lambda_2)q(x)$, where $q(x)$ has degree* 2. *There exists a field $\mathbb{Q}(\lambda)(\lambda_2)(\lambda_3)$, where λ_3 is a root of $q(x)$, and over the field $\mathbb{Q}(\lambda)(\lambda_2(\lambda_3)$, $q(x)$ factors into $(x-\lambda_3)(x-\lambda_4)$. Thus,*

$$x^4-3=(x-\lambda)(x-\lambda_2)(x-\lambda_3)(x-\lambda_4)$$

over $\mathbb{Q}(\lambda)(\lambda_2)(\lambda_3)(\lambda_4)$. In this particular example, we can take $\lambda_2=-\lambda$ and so $\mathbb{Q}(\lambda)=\mathbb{Q}(\lambda)(\lambda_2)$. Hence,

$$\mathbb{Q}(\lambda,\lambda_2,\lambda_3,\lambda_4)=\mathbb{Q}(\lambda,\lambda_3).$$

Now over $\mathbb{Q}(\lambda)$,

$$x^4-3=(x-\lambda)(x+\lambda)(x^2+\lambda^2).$$

Also, $x^2+\lambda^2$ is irreducible over $\mathbb{Q}(\lambda)$, a fact we leave as an exercise. Thus, $[\mathbb{Q}(\lambda):\mathbb{Q}]=4$ and $[\mathbb{Q}(\lambda)(\lambda_3):\mathbb{Q}(\lambda)]=2$. Hence, $[\mathbb{Q}(\lambda)(\lambda_3):\mathbb{Q}]=8$.

Example 16.2.3 leads us to believe that given any polynomial $f(x)$ in a polynomial ring $K[x]$ over a field K, there exists a field extension F/K such that $f(x)$ factors completely into linear factors. This is indeed the case, as we will presently show.

Definition 16.2.4 *Let K be a field. A polynomial $f(x)$ in $K[x]$ is said to* ***split*** *over a field $S\supseteq K$ if $f(x)$ can be factored as a product of linear factors in $S[x]$. A field S containing K is said to be a* ***splitting field*** *for $f(x)$ over K if $f(x)$ splits over S, but over no proper intermediate field of S/K.*

Example 16.2.5 *The field of complex numbers $\mathbb{C}$ is a splitting field for the polynomial x^2+1 over $\mathbb{R}$. This follows since $x^2+1=(x+i)(x-i)$ in $\mathbb{C}[x]$ and $\mathbb{C}/\mathbb{R}$ has no proper intermediate fields because $[\mathbb{C}:\mathbb{R}]=2$. (If $\mathbb{C}\supseteq L\supseteq\mathbb{R}$, where L is an intermediate field of $\mathbb{C}/\mathbb{R}$, then $2=[\mathbb{C}:L][L:\mathbb{R}]$ and so either $[\mathbb{C}:L]=1$ or $[L:\mathbb{R}]=1$. Thus, either $\mathbb{C}=L$ or $L=\mathbb{R}$.) Note that $\mathbb{C}$ is not the splitting field of x^2+1 over $\mathbb{Q}$ since x^2+1 splits over $\mathbb{Q}(i)\subset\mathbb{C}$.*

Theorem 16.2.6 *Let K be a field and $f(x)$ be a polynomial in $K[x]$ of degree n. Let F/K be a field extension. If*

$$f(x)=c(x-c_1)(x-c_2)\cdots(x-c_n) \text{ in } F[x],$$

then $K(c_1,c_2,\ldots,c_n)$ is a splitting field for $f(x)$ over K.

Proof. Since $c_1,c_2,\ldots,c_n$ are the roots of $f(x)$, $f(x)$ splits over $K(c_1,c_2,\ldots,c_n)$. Let L be an intermediate field of $K(c_1,c_2,\ldots,c_n)/K$ such that $f(x)$ splits over L. Since $K[x]$ is a UFD, there is only one way $f(x)$ can split over L, namely, $f(x)=c(x-c_1)(x-c_2)\cdots(x-c_n)$. Thus, $c_1,c_2,\ldots,c_n\in L$, whence $L\supseteq K(c_1,c_2,\ldots,c_n)$. Hence, $K(c_1,c_2,\ldots,c_n)$ is the smallest intermediate field over which $f(x)$ splits.

■

The field $\mathbb{Q}(\lambda,\ \lambda_3)$ of Example 16.2.3 is a splitting field for x^4-3 over $\mathbb{Q}$. We now prove the existence of splitting fields.

Theorem 16.2.7 *Let K be a field and $f(x)$ be a nonconstant polynomial over K. Then there is a splitting field for $f(x)$ over K.*

Proof. If $\deg f(x)=1$, then K is a splitting field for $f(x)$ over K. Assume the theorem is true for all polynomials of degree $n-1$ (≥ 1). Suppose $\deg f(x)=n$. There exists a field $K_1\supseteq K$ such that K_1 contains a root c_1 of $f(x)$ by Theorem 16.2.1. Thus, $f(x)=(x-c_1)f_1(x)$ in $K_1[x]$ and $\deg f_1(x)=n-1$. By the induction hypothesis, there exists a field extension E/K_1 such that $f_1(x)$ splits in $E[x]$. Thus, $f(x)$ splits in $E[x]$, say,

$$f(x)=c(x-c_1)(x-c_2)\cdots(x-c_n).$$

By Theorem 16.2.6, the intermediate field $K(c_1,\ c_2,\ \ldots,\ c_n)$ of E/K is a splitting field for $f(x)$ over K.

■

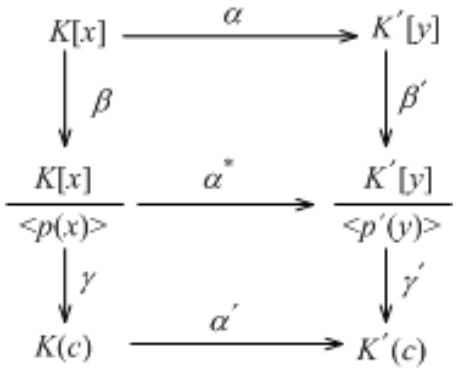
The intermediate field $\mathbb{Q}(\sqrt[4]{3},\ i\sqrt[4]{3})$ of $\mathbb{C}/\mathbb{Q}$ is a splitting field for x^4-3 over $\mathbb{Q}$. The field $\mathbb{Q}(\lambda,\ \lambda_3)$ of Example 16.2.3 is also a splitting field for x^4-3 over $\mathbb{Q}$. However, we cannot conclude that $\mathbb{Q}(\sqrt[4]{3},\ i\sqrt[4]{3})=\mathbb{Q}(\lambda,\ \lambda_3)$. Hence, splitting fields for a given polynomial over a field are not unique. We will show, however, that they are unique up to isomorphism.

Theorem 16.2.8 *Let α be an isomorphism of the field K onto the field K'. Let $p(x) = k_0+k_1x+k_2x^2+\cdots+k_nx^n$ be an irreducible polynomial in $K[x]$ of degree n, c be a root of $p(x)$ in some field extension of K, and $p'(y) = \alpha(k_0) + \alpha(k_1)y + \alpha(k_2)y^2 + \cdots + \alpha(k_n)y^n$ be the corresponding polynomial in $K'[y]$. Then $p'(y)$ is irreducible in $K'[y]$. If c' is a root of $p'(y)$ in some field extension of K', then α can be extended to an isomorphism α' of $K(c)$ onto $K'(c')$ with $\alpha'(c) = c'$. α' is the only extension of α such that $\alpha'(c) = c'$.*

Proof. By an argument similar to the one used in the proof of Theorem 11.1.14, α can be uniquely extended to an isomorphism $\overline{\alpha}$ of $K[x]$ onto $K'[y]$ so that for every polynomial $b_0 + b_1x + b_2x^2 + \cdots + b_mx^m \in K[x]$,

$$\overline{\alpha}(b_0 + b_1x + b_2x^2 + \cdots + b_mx^m) = \alpha(b_0) + \alpha(b_1)y + \alpha(b_2)y^2 + \cdots + \alpha(b_m)y^m.$$

We leave to the reader the verification that $p'(y)$ is irreducible in $K'[y]$. Let β be the natural homomorphisms of $K[x]$ onto $K[x]/\langle p(x)\rangle$ and β' be the natural homomorphisms of $K'[y]$ onto $K'[y]/\langle p'(y)\rangle$. Then Ker β = Ker $\beta' \circ \overline{\alpha}$. Hence, there exists an isomorphism α^* of $K[x]/\langle p(x)\rangle$ onto $K'[y]/\langle p'(y)\rangle$ such that $\beta' \circ \overline{\alpha} = \alpha^* \circ \beta$. By Theorem 16.1.11 and Corollary 16.1.12, there exist isomorphisms γ and γ' of $K[x]/\langle p(x)\rangle$ onto $K(c)$ and $K'[y]/\langle p'(y)\rangle$ onto $K'(c')$, respectively. Thus, α' is the map $\gamma' \circ \alpha^* \circ \gamma^{-1}$. The situation is described by the following diagram:



Let α'' be any other extension of α to an isomorphism of $K(c)$ onto $K'(c')$ such that $\alpha''(c) = c'$. Now $\{1, c, \ldots, c^{n-1}\}$ is a basis for $K(c)/K$ and $\{1, c', \ldots, c'^{n-1}\}$ is a basis for $K'(c')/K'$. We have that

$$\alpha''(\sum_{i=0}^{n-1} k_ic^i) = \sum_{i=0}^{n-1} \alpha''(k_i)\alpha''(c^i) = \sum_{i=0}^{n-1} \alpha(k_i)c'^i = \alpha'(\sum_{i=0}^{n-1} k_ic^i).$$

Hence, $\alpha'' = \alpha'$. ■

Corollary 16.2.9 *Let E/K be a field extension and $p(x)$ be an irreducible polynomial in $K[x]$. If $a, b \in E$ are roots of $p(x)$, then $K(a) \simeq K(b)$.*

Proof. Let $K = K'$ and α be the identity map.

■

From Corollary 16.2.9, we have $\mathbb{Q}(\sqrt[4]{3}) \simeq \mathbb{Q}(i\sqrt[4]{3})$ in Example 16.2.3.

Theorem 16.2.10 *Let α be an isomorphism from the field K onto the field K'. Let*

$$f(x) = k_0 + k_1x + k_2x^2 + \cdots + k_nx^n$$

be a polynomial in $K[x]$ and

$$f'(y) = \alpha(k_0) + \alpha(k_1)y + \alpha(k_2)y^2 + \cdots + \alpha(k_n)y^n$$

be the corresponding polynomial in $K'[y]$.

If S is a splitting field for $f(x)$ over K and S' is a splitting field for $f'(y)$ over K', then α can be extended to an isomorphism α' of S onto S'.

Proof. The proof is by induction on $\deg f(x)$. If $\deg f(x) = 1$, then $K = S$ and $K' = S'$. In this case, we can take $\alpha' = \alpha$. Assume the theorem is true for all polynomials of degree less than n (the induction hypothesis). Suppose $\deg f(x) = n$. Extend α to an isomorphism $\overline{\alpha}$ of $K[x]$ onto $K'[y]$ as in Theorem 16.2.8. Let $p(x)$ be an irreducible factor of $f(x)$ and $c_1 \in S$ be a root of $p(x)$. Let $c_1' \in S'$ be a root of $\overline{\alpha}(p(x)) = p(y)$. Then by Theorem 16.2.8, α can be extended to an isomorphism α_1 of $K(c_1)$ onto $K'(c_1')$. Extend α_1 to an isomorphism $\overline{\alpha_1}$ of $K(c_1)[x]$ onto $K'(c_1')[y]$. Now $f(x) = (x - c_1)f_1(x)$ in $K(c_1)[x]$ and $f'(y) = (y - c_1')f_1'(y)$ in $K(c_1')[y]$, where $f_1'(y) = \overline{\alpha_1}(f_1(x))$. Clearly S is a splitting field for $f_1(x)$ over $K(c_1)$ and S' is a splitting field for $f_1'(y)$ over $K'(c_1')$. Since $\deg f_1(x) = n - 1 = \deg f_1'(y)$, α_1 can be extended to an isomorphism of S onto S' by the induction hypothesis. ■

Corollary 16.2.11 *Let $f(x) \in K[x]$. Any two splitting fields for $f(x)$ over K are isomorphic.*

Proof. Let S and S' be two splitting fields for $f(x)$ over K. In Theorem 16.2.10, take $K = K'$ and α the identity mapping on K. ■

Definition 16.2.12 *Let F/K be a field extension and a, $b \in F$. Then a and b are called **conjugates** if a and b are roots of the same irreducible polynomial over K.*

We ask the reader to prove that the notion of conjugates defines an equivalence relation on F.

Example 16.2.13 *Consider the field extension $\mathbb{C}/\mathbb{R}$. Let $a, b \in \mathbb{R}$. Then $a + bi$ and its complex conjugate $a - bi$ are conjugates in the sense of Definition* 16.2.12. *This is obvious if $b = 0$. Suppose $b \neq 0$. Then $a + bi \notin \mathbb{R}$. Let $f(x) = x^2 - 2ax + (a^2 + b^2)$. Since $a + bi \notin \mathbb{R}$, $[\mathbb{R}(a + bi) : \mathbb{R}] = 2$. Now $a + bi$ is a root of $f(x)$ and $f(x)$ must be irreducible over $\mathbb{R}$. $a - bi$ is also a root of $f(x)$.*

In certain cases, the following theorem is useful in determining the irreducibility of a polynomial.

Theorem 16.2.14 *Let F be a field. Let p be a prime in $\mathbb{Z}$ and $a \in F$. Then the polynomial $x^p - a$ is reducible over F if and only if $x^p - a$ has a root in F.*

Proof. Suppose $f(x) = x^p - a \in F[x]$ is reducible. Let $f(x) = g(x)h(x)$ for some $g(x)$, $h(x) \in F[x]$, $\deg g(x) = m$, $0 < m < p$, and $0 < \deg h(x) < p$. Since $f(x)$ is monic, we can take $g(x)$ to be monic. By factoring $g(x)$ as a product of linear factors in a splitting field of $g(x)$ over F, we see that the constant term of $g(x)$ is $(-1)^m d$ for some $d \in F$. Since $\gcd(m, p) = 1$, there exist integers s and t such that $1 = sm + tp$. By Theorem 16.2.1, there is a field extension of F which contains a root of $f(x)$. Let b be such a root of $f(x)$.

Case 1: Suppose the characteristic of F is p. Since b is a root of $f(x)$, $b^p = a$. Thus,

$$(x - b)^p = x^p - b^p = x^p - a$$

and all the roots of $f(x)$ equal b. Now every root of $g(x)$ is also a root of $f(x)$. Thus, all the m roots of $g(x)$ are equal to b. Hence, $b^m = d$. Now

$$d^s = b^{ms} = b^{1-pt} = bb^{-pt} = ba^{-t}.$$

Hence, $b = d^s a^t \in F$ and so $f(x)$ has a root in F.

Case 2: Suppose that F is not of characteristic p. Let c be any other root of $f(x)$. Then

$$c^p = a = b^p.$$

Hence, $c = bu$, where $u = c^{-p+1}b^{p-1}$ and $u^p = 1$. From this, it follows that the roots of $f(x)$ are of the form

$$b, bu_1, \ldots, bu_{p-1},$$

where $u_i^p = 1$. As in case 1, we have that the product of the roots of $g(x)$ is

$$d = b^m u_1 u_2 \cdots u_{m-1} = b^m v,$$

where $v^p = 1$. Now $1 = sm + tp$ implies that

$$b^{sm} = v^{-s} d^s = b^{1-tp} = ba^{-t}.$$

Therefore, $b = v^{-s} d^s a^t$. It then follows that

$$a = b^p = (v^{-s} d^s a^t)^p = v^{-sp} (d^s a^t)^p = (d^s a^t)^p.$$

Thus, $d^s a^t \in F$ is a root of $f(x)$.

The converse follows from Corollary 11.1.10. ■

Worked-Out Exercises

♢ **Exercise 1**: Find a splitting field S of $x^4 - 10x^2 + 21$ over $\mathbb{Q}$. Find $[S : \mathbb{Q}]$ and a basis for $S/\mathbb{Q}$.

Solution: Note that $x^4 - 10x^2 + 21 = (x^2 - 3)(x^2 - 7)$ over $\mathbb{Q}$. Therefore, a splitting field S of $x^4 - 10x^2 + 21$ over $\mathbb{Q}$ is $\mathbb{Q}(\sqrt{3}, \sqrt{7})$. Hence, $[S : \mathbb{Q}] = 4$ and $\{1, \sqrt{3}, \sqrt{7}, \sqrt{21}\}$ is a basis for $S/\mathbb{Q}$, as can be seen from Worked-Out Exercise 2 (page 235).

♢ **Exercise 2**: Show that the splitting field of $x^p - 1$ over $\mathbb{Q}$ is of degree $p - 1$, where p is a prime.

Solution: Let $f(x) = x^p - 1 \in \mathbb{Q}[x]$. Now $f(x) = (x - 1)g(x)$, where $g(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$. Also,

$$g(x) = \frac{x^p - 1}{x - 1}.$$

Hence,

$$g(x + 1) = \frac{(x + 1)^p - 1}{x} = x^{p-1} + \binom{p}{1} x^{p-2} + \cdots + \binom{p}{p-1}.$$

Now since p is prime, $p | \binom{p}{r}$ for all $1 \leq r \leq p - 1$. Also, p^2 does not divide $\binom{p}{p-1}$. Therefore, by Eisenstein's criterion, $g(x + 1)$ is irreducible over $\mathbb{Q}$. Thus, $g(x)$ is irreducible over $\mathbb{Q}$. Let $\xi = e^{\frac{2\pi i}{p}}$, where $i^2 = -1$. Then the roots of $f(x)$ are $1, \xi, \xi^2, \ldots, \xi^{p-1}$ and the roots of $g(x)$ are $\xi, \xi^2, \ldots, \xi^{p-1}$. Now the splitting field of $f(x)$ is $S = \mathbb{Q}(1, \xi, \xi^2, \ldots, \xi^{p-1}) = \mathbb{Q}(\xi)$. Also, $g(x)$ is the minimal polynomial of ξ over $\mathbb{Q}$. Hence, $[S : \mathbb{Q}] = p - 1$.

♢ **Exercise 3**: Find the splitting field of the following polynomials over $\mathbb{Q}$.

(a) $x^4 + 1$.

(b) $x^6 + x^3 + 1$.

Solution:

(a) Let $f(x) = x^4 + 1$. Then $f(x) = (x^2 + \sqrt{2}x + 1)(x^2 - \sqrt{2}x + 1)$ over $\mathbb{Q}(\sqrt{2})$. Therefore, the roots of $f(x)$ are

$$\frac{-\sqrt{2} \pm i\sqrt{2}}{2}, \ \frac{\sqrt{2} \pm i\sqrt{2}}{2}.$$

Let S be the splitting field of $f(x)$ over $\mathbb{Q}$. We claim that $S = \mathbb{Q}(\sqrt{2}, i)$. Now

$$\sqrt{2} = \frac{\sqrt{2} + i\sqrt{2}}{2} + \frac{\sqrt{2} - i\sqrt{2}}{2} \in S$$

and

$$\sqrt{2}i = \frac{\sqrt{2} + i\sqrt{2}}{2} - \frac{\sqrt{2} - i\sqrt{2}}{2} \in S.$$

This implies that $i = \frac{i\sqrt{2}}{\sqrt{2}} \in S$. It now follows that $\mathbb{Q}(\sqrt{2}, i) \subseteq S$. Clearly $S \subseteq \mathbb{Q}(\sqrt{2}, i)$. Consequently, $S = \mathbb{Q}(\sqrt{2}, i)$. Now $x^2 - 2$ is the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$ and $x^2 + 1$ is the minimal polynomial of i over $\mathbb{Q}$. In fact, $x^2 + 1$ is the minimal polynomial of i over $\mathbb{Q}(\sqrt{2})$. Thus, $[S : \mathbb{Q}] = [S : \mathbb{Q}(\sqrt{2})][\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2 \cdot 2 = 4$.

(b) Let $f(x) = x^6 + x^3 + 1$. Now $(x^9 - 1) = (x^3 - 1)(x^6 + x^3 + 1)$. The roots of $(x^9 - 1)$ are $1, \xi, \xi^2, \ldots, \xi^8$ and $1, \xi^3, \xi^6$are the roots of $(x^3 - 1)$, where $\xi = e^{\frac{2\pi i}{9}}$. Hence, ξ, ξ^2, ξ^4, ξ^5, ξ^7, ξ^8 are the roots of $x^6 + x^3 + 1$. Therefore, $S = \mathbb{Q}(\xi, \xi^2, \xi^4, \xi^5, \xi^7, \xi^8) = \mathbb{Q}(\xi)$ is the splitting field of $x^6 + x^3 + 1$ over $\mathbb{Q}$. Since $x^6 + x^3 + 1$ is irreducible over $\mathbb{Q}$, $[S : \mathbb{Q}] = 6$.

Exercises

1. Prove that the polynomial $p'(y)$ in Theorem 16.2.8 is irreducible in $K'[y]$.
2. Let F/K be an algebraic field extension. Define $\sim$ on F by for all $a, b \in F$, $a \sim b$ if and only if a and b are conjugates. Prove that $\sim$ is an equivalence relation.
3. (i) Show that the polynomials $x^2 - 2x - 1$ and $x^2 - 2$ have the same splitting field over $\mathbb{Q}$.
   (ii) Find a pair of polynomials in $\mathbb{Q}[x]$, other than the pair given in (i), which have the same splitting field over $\mathbb{Q}$.
4. Find a splitting field S of the polynomial $x^3 - 3$ over $\mathbb{Q}$. Find $[S : \mathbb{Q}]$ and a basis for $S/\mathbb{Q}$.

5. Find a splitting field S of the polynomial $x^2 + x + [1]$ over $\mathbb{Z}_5$. Find $[S : \mathbb{Z}_5]$ and a basis for $S/\mathbb{Z}_5$.
6. Find a splitting field S of the polynomial $x^2 + [1]$ over $\mathbb{Z}_2$. Find $[S : \mathbb{Z}_2]$ and a basis for $S/\mathbb{Z}_2$.
7. Find a splitting field S of the polynomial $x^4 - 7x^2 + 10$ over $\mathbb{Q}$. Find $[S : \mathbb{Q}]$ and a basis for $S/\mathbb{Q}$.
8. Prove that $\mathbb{Q}(-\frac{1}{2} + \frac{\sqrt{3}}{2}i)$ is a splitting field of the polynomial $x^4 + x^2 + 1$ over $\mathbb{Q}$. Find $[\mathbb{Q}(-\frac{1}{2} + \frac{\sqrt{3}}{2}i) : \mathbb{Q}]$.
9. Let $f(x) \in K[x]$, a polynomial ring over the field K. Let S be a splitting field for $f(x)$ over K. Prove that for any field L, $S \supseteq L \supseteq K$, S is a splitting field of $f(x)$ over L.
10. Let $f(x), g(x)$, and $h(x) \in K[x]$, a polynomial ring over the field K. Suppose that S is a splitting field of $f(x)$ over K and $f(x) = g(x)h(x)$. Prove that S contains a splitting field of $g(x)$ over K.
11. Let $f(x), g(x) \in K[x]$, a polynomial ring over the field K. Suppose that $g(x) = f(ax+b)$, where $0 \neq a, b \in K$. Prove that $f(x)$ and $g(x)$ have equal splitting fields over K.
12. Prove that if $f(x)$ is a polynomial in $K[x]$ of degree n, then $[S : K] \leq n!$, where S is a splitting field of $f(x)$ over K.
13. Let K be a field and $f_1(x)$, $f_2(x), \ldots, f_n(x) \in K[x]$ be such that $\deg f_i(x) \geq 1$, $1 \leq i \leq n$. Show that there exists a field extension F/K such that each $f_i(x)$ has a root in F.
14. Let F be a field of prime characteristic p and $a \in F$. Prove that $x^p - x - a$ is reducible over F if and only if $x^p - x - a$ has a root in F.
15. Answer the following statements, true or false. If the statement is true, prove it. If it is false, give a counter example.

    (i) Let $f(x)$ be an irreducible polynomial of degree n over a field K of characteristic 0. Let $S = K(c_1, c_2, \ldots, c_n)$ be a splitting field of $f(x)$ over K, where $c_1, c_2, \ldots, c_n$ are the roots of $f(x)$. Then $K(c_2, \ldots, c_n) \subset S$.

    (ii) The polynomial $f(x) = x^5 - x - 30$ is reducible over $\mathbb{Q}$.

    (iii) $\mathbb{C}$ is a splitting field of some polynomial over $\mathbb{Q}$.

16.3 Algebraically Closed Fields

The most important result in Steinitz's work in 1910 was his proof of the existence and uniqueness of an algebraic closure of a field. In this section[1], we present these results.

Definition 16.3.1 *A field K is called* ***algebraically closed*** *if for all $f(x) \in K[x]$ with $\deg f(x) \geq 1$, $f(x)$ has a root in K.*

Theorem 16.3.2 *Let K be a field. The following conditions are equivalent.*

(i) K is algebraically closed.

(ii) Every irreducible polynomial in $K[x]$ is of degree 1.

(iii) Let $f(x) \in K[x]$, $\deg f(x) \geq 1$. Then $f(x)$ splits as a product of linear factors over K.

(iv) If F/K is an algebraic field extension, then $F = K$.

Proof. (i)⇒(ii) Let $p(x) \in K[x]$ and $p(x)$ be irreducible. By (i), there exists $a \in K$ such that $p(a) = 0$. Then $p(x) = (x - a)g(x)$ for some $g(x) \in K[x]$. Since $p(x)$ is irreducible, $g(x) \in K$. Hence, $\deg p(x) = 1$.

(ii)⇒(iii) Let $f(x) \in K[x]$ and $\deg f(x) \geq 1$. Let $f(x) = p_1(x) \cdots p_s(x)$, where $p_i(x) \in K[x]$ is irreducible, $1 \leq i \leq s$. Then $\deg p_i(x) = 1$, $1 \leq i \leq s$. We may write $p_i(x) = k_i(x - a_i)$, where $k_i, a_i \in K$, $1 \leq i \leq s$. Let $k = k_1 \cdots k_s$. Then $f(x) = k(x - a_1) \cdots (x - a_s)$. Thus, $f(x)$ splits as a product of linear factors over K.

(iii)⇒(iv) Let F/K be an algebraic field extension. Let $c \in F$ and let $p(x) \in K[x]$ be the minimal polynomial of c over K. Since $p(x)$ is irreducible, $\deg p(x) = 1$ by (iii). Therefore, $p(x) = ax + b \in K[x]$. Since $p(c) = 0$, $ac + b = 0$. Thus, $c = -a^{-1}b \in K$. Hence, $K = F$.

(iv)⇒(i) Let $f(x) \in K[x]$, $\deg f(x) \geq 1$. There exists a field extension F/K such that F has a root of $f(x)$, say, a. Then $K(a)/K$ is an algebraic field extension. Therefore, $K(a) = K$ and so $a \in K$. Thus, K is algebraically closed.

■

We now prove the existence of an algebraically closed field. The following proof is due to Artin.

[1] This section may be skipped without any discontinuity. The only place this section is needed is in Exercise 4 (Section 24.1).

Theorem 16.3.3 *Let K be a field. Then there exists an algebraically closed field F such that K is a subfield of F.*

Proof. We first construct an extension F_1/K such that if $f(x) \in K[x]$ and $\deg f(x) \geq 1$, then $f(x)$ has a root in F_1. Let $\mathcal{K}$ be the set of all polynomials in $K[x]$ of degree ≥ 1. Let S be a set which is in one-one correspondence with $\mathcal{K}$. For $f(x) \in \mathcal{K}$, let x_f be the corresponding element in S.

Consider the polynomial ring $K[S]$. Let I be the ideal of $K[S]$ generated by all polynomials $f(x_f)$ in $K[S]$. We claim that $I \neq K[S]$. Suppose that $I = K[S]$. Then there exists $g_i \in K[S]$ such that

$$g_1 f_1(x_{f_1}) + g_2 f_2(x_{f_2}) + \cdots + g_n f_n(x_{f_n}) = 1. \tag{16.7}$$

Write $x_i = x_{f_i}$, $1 \leq i \leq n$. Since the polynomials g_i, $1 \leq i \leq n$, involve only a finite number of indeterminates, say, $x_1, x_2, \ldots, x_m$, with $m \geq n$, we may write Eq. (16.7) as

$$\sum_{i=1}^{n} g_i(x_1, x_2, \ldots, x_m) f_i(x_i) = 1. \tag{16.8}$$

By Exercise 13 (page 243), there exists a finite extension L/K such that each polynomial f_i, $1 \leq i \leq n$, has a root in L. Let c_i be a root of f_i in L, $1 \leq i \leq n$. Let $c_i = 0$ for $n < i \leq m$. Substituting c_i for x_i, $1 \leq i \leq n$, in Eq. (16.8), we get $0 = 1$, a contradiction. Hence, $I \neq K[S]$.

Let M be a maximal ideal of $K[S]$ such that $I \subseteq M$. Let $F_1 = K[S]/M$. Then F_1 is a field containing an isomorphic copy $(K + M)/M$ of K. Thus, F_1 can be regarded as a field extension of K. Also, if $f \in K[x]$ and $\deg f(x) \geq 1$, then $x_f + M$ is a root of f in F_1.

By induction, we can form a chain of fields

$$F_1 \subseteq F_2 \subseteq \cdots \subseteq F_n \subseteq \cdots$$

such that every polynomial of degree ≥ 1 in F_n has a root in F_{n+1}. Let $F = \cup_{i=1}^{\infty} F_n$. Then F is a field. Let $f \in F[x]$. Then $f \in F_n[x]$ for some positive integer n. Thus f has a root in $F_{n+1} \subseteq F$. Hence, F is algebraically closed. ■

Corollary 16.3.4 *Let K be a field. Then there exists an algebraic field extension F/K such that F is algebraically closed.*

Proof. By Theorem 16.3.3, there exists a field extension E/K such that E is algebraically closed. Let $F = \{a \in E \mid a \text{ is algebraic over } K\}$. Then F/K is an algebraic extension. Let $f(x) \in F[x]$ and $\deg f(x) \geq 1$. Then $f(x)$ has a root c in E. Thus, c is algebraic over F. Since F/K is an algebraic extension, c is algebraic over K. Hence, $c \in F$ and so F is algebraically closed. ■

Definition 16.3.5 *Let K be a field. A field $F \supseteq K$ is called an* ***algebraic closure*** *of K if*
(i) F/K is algebraic and
(ii) F is algebraically closed.

For any field K, Corollary 16.3.4 guarantees the existence of an algebraic closure of K.

Lemma 16.3.6 *Let F and L be fields with L algebraically closed. Let $\sigma : F \to L$ be an isomorphism of F into L. Let a be an algebraic element over F in some field extension of F. Let $f(x) \in F[x]$ be the minimal polynomial of a. Then σ can be extended to an isomorphism η of $F(a)$ into L and the number of such extensions is equal to the number of distinct roots of $f(x)$.*

Proof. Let $f(x) = a_0 + a_1 x + \cdots + a_n x^n \in F[x]$ and $f^\sigma(x) = \sigma(a_0) + \sigma(a_1)x + \cdots + \sigma(a_n)x^n \in L[x]$. Since L is algebraically closed there exists a root b of $f^\sigma(x)$ in L. Since a is algebraic over F, $F(a) = F[a]$ by Corollary 16.1.12. Thus, if $u \in F(a)$, then $u = c_0 + c_1 a + \cdots + c_k a^k \in F[a]$. Define $\eta : F(a) \to L$ by

$$\eta(c_0 + c_1 a + \cdots + c_k a^k) = \sigma(c_0) + \sigma(c_1)b + \cdots + \sigma(c_k)b^k$$

for all $c_0 + c_1 a + \cdots + c_k a^k \in F(a)$. Suppose $c_0 + c_1 a + \cdots + c_k a^k = d_0 + d_1 a + \cdots + d_s a^s$. Let $\gamma(x) = c_0 + c_1 x + \cdots + c_k x^k$ and $\gamma'(x) = d_0 + d_1 x + \cdots + d_s x^s$. Then $(\gamma - \gamma')(a) = 0$. Hence, $f(x)$ divides $(\gamma - \gamma')(x)$. Thus, $f^\sigma(x)$ divides $(\gamma^\sigma - \gamma'^\sigma)(x)$. Consequently, $(\gamma^\sigma - \gamma'^\sigma)(b) = 0$ and so $\sigma(c_0) + \sigma(c_1)b + \cdots + \sigma(c_k)b^k = \sigma(d_0) + \sigma(d_1)b + \cdots + \sigma(d_s)b^s$. Thus, η is well defined. Clearly η is an isomorphism. The number of distinct roots of $f(x)$ in the algebraic closure of F is equal to the number of distinct roots of $f^\sigma(x)$ in L. For any extension $\xi : F(a) \to L, \xi(a)$ is a root of

$f^\sigma(x)$. Therefore, the number of such extensions is equal to the number of distinct roots of $f(x)$.
■

We close this section by showing that the algebraic closure of a field is unique up to isomorphism. Our proof uses Zorn's lemma while Steinitz's original proof used the equivalent concept of the axiom of choice.

Theorem 16.3.7 *Let F/K be an algebraic field extension. Let L be an algebraically closed field and σ be an isomorphism of K into L. Then there exists an isomorphism η of F into L such that $\eta|_K = \sigma$.*

Proof. Let $\mathcal{S} = \{(E, \lambda) \mid E$ is a subfield of F, $K \subseteq E$ and $\lambda : E \to L$ is an isomorphism such that $\lambda|_K = \sigma\}$. Since $(K, \sigma) \in \mathcal{S}$, $\mathcal{S} \neq \phi$. Let (E, λ), $(E', \lambda') \in \mathcal{S}$. Define a relation $\leq$ on $\mathcal{S}$ by $(E, \lambda) \leq (E', \lambda')$ if $E \subseteq E'$ and $\lambda'|_E = \lambda$. Then $(S, \leq)$ is a poset. Let $\{(E_i, \lambda_i)\}_{i\in\Lambda}$ be a chain in $\mathcal{S}$. Let $E = \cup_{i\in\Lambda} E_i$. Then E is a field and $K \subseteq E$. Define $\lambda : E \to L$ as follows: Let $a \in E$. Then $a \in E_n$ for some n. Define $\lambda(a) = \lambda_n(a)$. Since $\{(E_i, \lambda_i)\}_{i\in\Lambda}$ is a chain, λ is an isomorphism of E into L. Hence, $(E, \lambda) \in \mathcal{S}$ and (E, λ) is an upper bound of $\{(E_i, \lambda_i)\}_{i\in\Lambda}$. Hence, by Zorn's lemma, $\mathcal{S}$ has a maximal element, say, (T, η). Suppose $T \neq F$. Let $a \in F\backslash T$. By Lemma 16.3.6, there exists an isomorphism $\beta : T(a) \to L$ such that $\beta|_T = \eta$. From this, it follows that $(T(a), \beta) \in \mathcal{S}$, a contradiction of the maximality of (T, η). Thus, $F = T$. ■

Theorem 16.3.8 *Let K be a field. Let F and F' be two algebraic closures of K. Then there exists an isomorphism λ of F onto F' such that $\lambda(a) = a$ for all $a \in K$.*

Proof. Let $\sigma : K \to F'$ be such that $\sigma(a) = a$ for all $a \in K$. Then σ is an isomorphism of K into F'. By Theorem 16.3.7, there exists an isomorphism $\lambda : F \to F'$ such that $\lambda|_K = \sigma$. Now $\lambda(F) \simeq F$. Thus, $\lambda(F)$ is algebraically closed and $K \subseteq \lambda(F)$. Now $K \subseteq \lambda(F) \subseteq F'$. Since F'/K is algebraic, $F'/\lambda(F)$ is algebraic. Thus, $F' = \lambda(F)$. Hence, $F \simeq F'$. ■

Exercises

1. If F is a field with a finite number of elements, prove that F is not algebraically closed.

Chapter 17

Multiplicity of Roots

17.1 Multiplicity of Roots

In some cases, an irreducible polynomial $p(x)$ of degree n over a field K does not have n distinct roots in a splitting field of $p(x)$ over K. In this chapter, we examine this situation.

If $f(x)$ is a polynomial over K and c is a root of $f(x)$ in some field F containing K, then the **multiplicity** of c is the largest positive integer m such that $(x-c)^m$ divides $f(x)$ over F.

Definition 17.1.1 *Let K be a field and $p(x)$ be an irreducible polynomial in $K[x]$ of degree n. Then $p(x)$ is called **separable** if it has n distinct roots in a splitting field S of $p(x)$ over K; otherwise $p(x)$ is called **inseparable** over K. An arbitrary polynomial in $K[x]$ is called **separable** if each of its irreducible factors in $K[x]$ is separable; otherwise it is called **inseparable**.*

Definition 17.1.2 *Let F/K be a field extension and c be an element of F which is algebraic over K. Then c is called **separable** (or **separable algebraic**) over K if its minimal polynomial over K is separable; otherwise c is called **inseparable** over K. If F/K is an algebraic extension, then F/K is called **separable** (or **separable algebraic**) if every element of F is separable over K; otherwise F/K is called **inseparable**.*

Let F/K be a field extension and L be an intermediate field of F/K. Let $c \in F$ and suppose c is separable over K. Then c must be separable over L. This follows since if $f(x)$ and $p(x)$ are the minimal polynomials of c over K and L, respectively, then $p(x)|f(x)$. Hence, c cannot be a multiple root of $p(x)$ since it is not one of $f(x)$.

Example 17.1.3 *Consider the field $K(t)$, where K is a field of prime characteristic p and t is transcendental over K. It follows that the polynomial $x^p - t^p$ is irreducible over $K(t^p)$ by Eisenstein's criterion since t^p is irreducible in $K[t^p]$. Now $x^p - t^p$ factors into*

$$\underbrace{(x-t)(x-t)\cdots(x-t)}_{p \text{ times}} = (x-t)^p$$

over $K(t)$. Thus, $K(t)$ is a splitting field for $x^p - t^p$ over $K(t^p)$ and we see that $x^p - t^p$ has only one root in $K(t)$, namely, t. (Since $t \notin K(t^p)$, we can also use Theorem 16.2.14 to deduce that $x^p - t^p$ is irreducible over $K(t^p)$.) Thus, $x^p - t^p$, t, and $K(t)$ are inseparable over $K(t^p)$. Note that t has multiplicity p over $K(t^p)$.

Let K be a field and

$$f(x) = k_0 + k_1 x + \cdots + k_n x^n$$

be a polynomial in $K[x]$. Then by the **formal derivative**, $f'(x)$, of $f(x)$ we mean the polynomial

$$f'(x) = k_1 + \cdots + ik_i x^{i-1} + \cdots + nk_n x^{n-1} \in K[x].$$

Let K be a field and $f(x)$, $g(x) \in K[x]$. The following properties of formal derivatives are easily verified:

$$\begin{aligned} (f(x)+g(x))' &= f'(x)+g'(x), \\ (f(x)g(x))' &= f(x)g'(x)+f'(x)g(x), \\ (kf(x))' &= kf'(x) \text{ for all } k \in K \end{aligned}$$

and if $f(x) = x$, then $f'(x) = 1$.

Theorem 17.1.4 *Let K be a field and $f(x) \in K[x]$, $f(x) \neq 0$. Let a be a root of $f(x)$ in some extension field F of K. Then a is a multiple root of $f(x)$ if and only if $f'(a) = 0$.*

Proof. Suppose a is a multiple root of $f(x)$. Then $(x-a)^2$ divides $f(x)$. Hence,

$$f(x) = (x-a)^2 g(x)$$

for some $g(x) \in F[x]$. Now $f'(x) = (x-a)\{(x-a)g'(x) + 2g(x)\}$. Therefore, $f'(a) = 0$. Conversely, suppose $f'(a) = 0$. Then $\deg f(x) \geq 2$. By the division algorithm,

$$f(x) = (x-a)^2 q(x) + h(x)$$

for some $q(x), h(x) \in F[x]$, where either $h(x) = 0$ or $\deg h(x) \leq 1$. Suppose $h(x) \neq 0$. Since $f(a) = 0$, $h(a) = 0$. Thus, $\deg h(x) = 1$ and a is a root of $h(x)$. Hence, $h(x) = b(x-a)$ for some $0 \neq b \in K$. This implies that

$$f(x) = (x-a)^2 q(x) + b(x-a)$$

and so

$$f'(x) = (x-a)\{(x-a)q'(x) + 2q(x)\} + b$$

Therefore,

$$0 = f'(a) = b,$$

a contradiction. Hence, $h(x) = 0$ and so $f(x) = (x-a)^2 q(x)$. Consequently, a is a multiple root of $f(x)$. ■

Theorem 17.1.5 *For any field K, an irreducible polynomial $p(x)$ in $K[x]$ is separable if and only if $p(x)$ and its formal derivative $p'(x)$ are relatively prime.*

Proof. Let $d(x)$ denote the gcd of $p(x)$ and $p'(x)$. Suppose $p(x)$ is separable. Let c be a root of $p(x)$ in some field containing K. Then $p(x) = (x-c)f(x)$ for some $f(x) \in K(c)[x]$. Since $p(x)$ is separable, $f(c) \neq 0$. Now $p'(x) = f(x) + (x-c)f'(x)$ and so $p'(c) = f(c) + 0 \neq 0$. Hence, c is not a root of $d(x)$. But every root of $d(x)$ must be a root of $p(x)$ since $d(x) | p(x)$. Thus, since we have just seen that $d(x)$ and $p(x)$ have no common roots, $d(x)$ has no roots. Therefore, $d(x) = 1$.

Conversely, suppose that $d(x) = 1$. Let c be any root of $p(x)$. Let m denote the multiplicity of c. Then

$$p(x) = (x-c)^m f(x)$$

over $K(c)$ and c is not a root of $f(x)$. Now

$$\begin{aligned} p'(x) &= m(x-c)^{m-1} f(x) + (x-c)^m f'(x) \\ &= (x-c)^{m-1}[mf(x) + (x-c)f'(x)]. \end{aligned}$$

Thus, $(x-c)^{m-1}$ is a common divisor of $p'(x)$ and $p(x)$. Hence,

$$(x-c)^{m-1} | d(x).$$

Since $d(x) = 1$, $m = 1$. Consequently, $p(x)$ has no repeated roots. ■

Theorem 17.1.6 *For any field K, an irreducible polynomial $p(x)$ in $K[x]$ is separable if and only if $p'(x) \neq 0$.*

Proof. Let $d(x)$ denote the gcd of $p(x)$ and $p'(x)$. Suppose $p(x)$ is separable. If $p'(x) = 0$, then $d(x) = p(x) \neq 1$, a contradiction of Theorem 17.1.5. Conversely, suppose $p'(x) \neq 0$. Since $p(x)$ is irreducible, the only common divisors of $p(x)$ and $p'(x)$ are 1 and $p(x)$. Since $1 \leq \deg p'(x) < \deg p(x)$, 1 is the only common divisor of $p'(x)$ and $p(x)$. Hence, $d(x) = 1$. Thus, $p(x)$ is separable by Theorem 17.1.5. ■

Corollary 17.1.7 *Let K be a field of characteristic 0. Then every nonconstant polynomial in $K[x]$ is separable.*

Proof. Let $f(x)$ be any nonconstant polynomial in $K[x]$ and $p(x) = k_0 + k_1 x + k_2 x^2 + \cdots + k_n x^n$ be any irreducible factor of $f(x)$, where $n \geq 1$. Then there exists $i > 0$ such that $k_i \neq 0$. Hence, $ik_i \neq 0$ since K has characteristic 0. Thus, $p'(x) \neq 0$ and so $p(x)$ is separable by Theorem 17.1.6. Hence, $f(x)$ is separable. ■

Example 17.1.8 *Consider the irreducible polynomial $p(x) = x^p - t^p$ over $K(t^p)$ of Example* 17.1.3. *Then $p'(x) = px^{p-1} = 0$. Thus, $x^p - t^p$ is inseparable over $K(t^p)$.*

Theorem 17.1.9 *Let K be a field of characteristic $p > 0$. Then an irreducible polynomial $p(x) = k_0 + k_1x + k_2x^2 + \cdots + k_nx^n$ over K is inseparable if and only if $p(x) = q(x^p)$ for some $q(x^p) \in K[x^p]$.*

Proof. Clearly $p'(x) = 0$ if and only if $ik_i = 0$ for all $i = 1, 2, \ldots, n$. Thus, $p'(x) = 0$ if and only if $p|i$ for those i such that $k_i \neq 0$. Hence, $p'(x) = 0$ if and only if $p(x) = q(x^p)$ for some $q(x^p) \in K[x^p]$. The conclusion now follows from Theorem 17.1.6.
■

Let K be a field of characteristic $p > 0$. Let $K^p = \{a^p \mid a \in K\}$. The reader is asked to verify in Exercise 7 (page 256) that K^p is a subfield of K.

Definition 17.1.10 *Let K be a field. Then K is called* ***perfect*** *if every algebraic extension of K is separable.*

Example 17.1.11 *By Corollary* 17.1.7*, every field of characteristic* 0 *is perfect.*

The following theorem gives a necessary and sufficient condition for a field to be perfect.

Theorem 17.1.12 *Let K be a field of characteristic $p > 0$. Then K is perfect if and only if $K = K^p$.*

Proof. Suppose K is perfect. Let $a \in K$ and F be a splitting field of $x^p - a \in K[x]$. Then F/K is a separable extension. Let $b \in F$ be a root of $x^p - a$. Then

$$x^p - a = (x - b)^p.$$

Let $p(x) \in K[x]$ be the minimal polynomial of b. Then $p(x)$ has distinct roots. If $\deg p(x) > 1$, then since $p(x)|(x - b)^p$, $p(x)$ has multiple roots, a contradiction. Hence, $\deg p(x) = 1$. This implies that $b \in K$. Hence, $a = b^p \in K^p$. Thus, $K = K^p$.

Conversely, suppose $K = K^p$. Let F/K be an algebraic field extension. Let $a \in F$ and $f(x) \in K[x]$ be the minimal polynomial of a. Suppose $f(x)$ is not separable. Then by Theorem 17.1.9, $f(x) = g(x^p)$ for some $g(x) \in K[x]$. Hence,

$$f(x) = a_0 + a_1x^p + \cdots + a_kx^{pk},$$

$a_i \in K$, $1 \leq i \leq k$. Since $K = K^p$, $a_i = b_i^p$ for some $b_i \in K$, $1 \leq i \leq k$. Therefore,

$$f(x) = (b_0 + b_1x + \cdots + b_kx^k)^p,$$

a contradiction, since $f(x)$ is irreducible over K. Hence, $f(x)$ is separable. Thus, F/K is a separable extension. Consequently, K is perfect. ■

Example 17.1.13 *Let K be a finite field of characteristic p. Define $\sigma : K \to K^p$ by $\sigma(a) = a^p$. Then σ is a homomorphism. Suppose that $\sigma(a) = \sigma(b)$. Then $a^p = b^p$ and so $(a - b)^p = 0$. Since K is a field, K has no nonzero nilpotent elements. Thus, $a = b$ and so σ is one-one. Hence, $|K| = |\sigma(K)| \leq |K^p| \leq |K|$ and so $|K| = |K^p|$. Since K^p is a subfield of K and K is finite, $K = K^p$. Hence, K is perfect. We have thus shown that every finite field is perfect.*

If $p(x) = k_0 + k_1x + k_2x^2 + \cdots + k_nx^n$ is irreducible and inseparable over K in Theorem 17.1.9, then $p(x) = k_0 + k_px^p + \cdots + k_{pm}(x^p)^m = q(x^p)$. It may be the case that $p(x) = q(x^p) = s(x^{p^2})$ in $K[x^{p^2}]$. However, there exists a largest positive integer e such that $p(x) = t(x^{p^e})$ for some $t(x^{p^e}) \in K[x^{p^e}]$. If $n = \deg p(x)$, then $p^e|n$.

Definition 17.1.14 *Let K be a field of characteristic $p > 0$ and $p(x)$ be an irreducible polynomial in $K[x]$. Let e be the largest nonnegative integer such that $p(x) = q(x^{p^e})$ for some $q(x^{p^e}) \in K[x^{p^e}]$. Then e is called the* ***exponent of inseparability*** *of $p(x)$ and p^e is called the* ***degree of inseparability*** *of $p(x)$. If n denotes the degree of $p(x)$, then $n_0 = \frac{n}{p^e}$ is called the* ***degree of separability*** *or* ***reduced degree*** *of $p(x)$ over K.*

By Theorem 17.1.9, $p(x)$ in Definition 17.1.14 is separable if and only if $e = 0$.

Theorem 17.1.15 *Let K be a field of characteristic $p > 0$ and*

$$p(x) = k_{n_0}(x^{p^e})^{n_0} + \cdots + k_1x^{p^e} + k_0$$

be an irreducible polynomial in $K[x]$, where e is the exponent of inseparability of $p(x)$. Then the polynomial

$$s(y) = k_{n_0}y^{n_0} + \cdots + k_1y + k_0 \in K[y]$$

is irreducible and separable over K.

Proof. If $s(y) = f(y)g(y) \in K[y]$, then $p(x) = f(x^{p^e})g(x^{p^e})$, contrary to the fact that $p(x)$ is irreducible in $K[x]$. Thus, $s(y)$ is irreducible in $K[y]$. If $s(y) = q(y^p)$ for some $q(y^p) \in K[y^p]$, then $p(x) = q(x^{p^{e+1}})$, contrary to the maximality of e. Hence, $s(y)$ is separable. ■

Example 17.1.16 *Consider the polynomial $p(x) = x^{2p} + tx^p + t$ over the field $K(t)$, where K is a field of characteristic $p > 0$ and t is transcendental over K. By Eisenstein's criterion, $p(x)$ is irreducible over $K(t)$. Now $p(x) = (x^p)^2 + tx^p + t \in K(t)[x]$ and so $p(x)$ is inseparable over $K(t)$. The inseparability exponent e of $p(x)$ equals 1. Thus, $x^2 + tx + t$ is separable over $K(t)$.*

Definition 17.1.17 *Let F/K be a field extension. F is called a **simple** extension if $F = K(a)$ for some $a \in F$. Such an element a is called a **primitive** element.*

Theorem 17.1.18 *Let K be an infinite field and $K(a, b)/K$ be a field extension with a algebraic over K and b separable algebraic over K. Then there exists an element $c \in K(a, b)$ such that $K(a, b) = K(c)$, i.e., $K(a, b)/K$ is a simple extension.*

Proof. Let $f(x)$ and $g(x)$ be the minimal polynomials of a and b over K with degrees n and m and roots $a = a_1, a_2, \ldots, a_n$, and $b = b_1, b_2, \ldots, b_m$, respectively, in some extension field of K. Since b is separable, all b_i's are distinct. Also, since K is infinite, there exists $s \in K$ such that $a + sb \neq a_i + sb_j$, i.e.,

$$s \neq \frac{a_i - a}{b - b_j}$$

for all $1 \leq i \leq n$, $1 < j \leq m$. Let $c = a + sb$. Then $c - sb_j \neq a_i$ for all $1 \leq i \leq n$, $1 < j \leq m$. Also, $K(c) \subseteq K(a, b)$. Let $h(x) = f(c - sx) \in K(c)[x]$. Now

$$h(b) = f(c - sb) = f(a) = 0.$$

Thus, $g(x)$ and $h(x)$ have the common root b of multiplicity 1 in the field $K(a, b)$. Now

$$h(b_j) = f(c - sb_j) \neq 0$$

for all $1 < j \leq m$. Thus, $g(x)$ and $h(x)$ have only root b in common. Let $d(x) \in K(c)[x]$ be the greatest common divisor of $g(x)$ and $h(x)$. Then b is a root of $d(x)$. Every root of $d(x)$ is also a root of $g(x)$ and $h(x)$. Since $g(x)$ and $h(x)$ have no roots other than b in common in any field and b is of multiplicity 1, $d(x)$ is of degree 1. Hence, $d(x) = x - b$. But then $b \in K(c)$. Thus, $a = c - sb \in K(c)$. Therefore, $K(a, b) \subseteq K(c) \subseteq K(a, b)$ and so $K(c) = K(a, b)$. ■

Corollary 17.1.19 *Let K be an infinite field. Let $a_1, a_2, \ldots, a_n$ be elements in some field containing K. Suppose that a_1 is algebraic and $a_2, \ldots, a_n$ are separable algebraic over K. Then there exists an element $c \in K(a_1, \ldots, a_n)$ such that $K(c) = K(a_1, \ldots, a_n)$, i.e., $K(a_1, \ldots, a_n)/K$ is a simple extension.*

Proof. The result follows by induction on n and Theorem 17.1.18. ■

Corollary 17.1.20 *Let F/K be a field extension and the characteristic of K be 0. Let $a_1, a_2, \ldots, a_n \in F$ be algebraic over K. Then $K(a_1, \ldots, a_n)/K$ is a simple extension.*

Proof. The proof follows by Corollaries 17.1.7 and 17.1.19. ■

Example 17.1.21 *Consider $\mathbb{Q}(\sqrt{2}, i)$. Now $1 \neq \frac{-\sqrt{2}-(\sqrt{2})}{i-(-i)} = -\frac{\sqrt{2}}{i}$. Thus, $\mathbb{Q}(\sqrt{2}, i) = \mathbb{Q}(\sqrt{2} + i)$ by the proof of Theorem* 17.1.18, *with $s = 1$ there.*

Theorem 17.1.22 (Artin) *Let K be an infinite field. Let F/K be a finite field extension. Then F/K is a simple extension if and only if there are only a finite number of intermediate fields of F/K.*

Proof. Suppose F/K is a simple extension. Let $F = K(a)$ for some $a \in F$. Let L be an intermediate field of F/K and $f(x)$ be the minimal polynomial of a over L. Let L' be the field generated by K and the coefficients of $f(x)$. Then $L' \subseteq L$ and $f(x)$ is also the minimal polynomial of a over L'. Hence,

$$[F : L] = \deg f(x) = [F : L'].$$

Thus, $[L : L'] = 1$ and so $L = L'$. Let $g(x)$ be the minimal polynomial of a over K. Then $f(x)$ divides $g(x)$. Now $g(x)$ has only a finite number of distinct monic factors. Hence, the number of intermediate fields is finite.

Conversely, suppose there are only a finite number of intermediate fields of F/K. Let $a, b \in F$. We first show that $K(a,b)/K$ is a simple extension. Let $c \in K$ and $F_c = K(a+cb)$. Then for all $c \in K$, F_c is an intermediate field of $K(a,b)/K$. Since the number of intermediate fields is finite and K is infinite, there exists $c, d \in K$, $c \neq d$ such that $F_c = F_d$. Then

$$b = (c-d)^{-1}(a+cb-a-db) \in F_c.$$

Hence, $a = a + cb - cb \in F_c$. Thus, $K(a,b) = F_c = K(a+cb)$, i.e., $K(a,b)/K$ is a simple extension. Now for all $a \in F$, $K(a)$ is an intermediate field of F/K. Since $[F:K]$ is finite, $[K(a):K]$ is finite. Let

$$A = \{[K(a):K] \mid a \in F\}.$$

Then A is a finite subset of $\mathbb{Z}$. Let $a \in F$ be such that the maximum of $A = [K(a):K]$. Suppose $F \neq K(a)$. Let $b \in F$ be such that $b \notin K(a)$. Then $K(a) \subset K(a,b)$. There exists $c \in F$ such that $K(a,b) = K(c)$. Therefore, $K(a) \subset K(c)$. Hence, $[K(c):K] > [K(a):K]$, a contradiction to the maximality of $[K(a):K]$. Consequently, $F = K(a)$, i.e., F/K is a simple extension. ■

Let F/K be a field extension. In the next chapter, we show that every finite extension of a finite field is a simple extension (Corollary 18.1.8, page 258). Hence, from this and Theorem 17.1.22, it follows that F/K is a simple extension if and only if there are only a finite number of intermediate fields of F/K.

We now focus our attention on the study of **separable algebraic and purely inseparable extensions**.[1]

Theorem 17.1.23 *Let K be a field of characteristic $p > 0$ and $f(x) = x^{p^e} - k$ be a polynomial over K, where e is a positive integer. Then $f(x)$ is irreducible over K if and only if $k \notin K^p$.*

Proof. Suppose $f(x)$ is irreducible over K. If $k = k'^p \in K^p$ for some $k' \in K$, then $f(x) = (x^{p^{e-1}} - k')^p$, contrary to the fact that $f(x)$ is irreducible over K. Hence, $k \notin K^p$. Conversely, suppose $k \notin K^p$. Let $p(x)$ be a nonconstant monic irreducible factor of $f(x)$ in $K[x]$ and c be a root of $p(x)$. Then c is a root of $f(x)$ and so $c^{p^e} = k$ and $f(x) = (x-c)^{p^e}$ over $K(c)$. Since $K(c)[x]$ is a unique factorization domain, it follows that $p(x)$ is some power of $(x-c)$, say, $p(x) = (x-c)^m$. Thus, $mn = p^e$ for some n so that $m = p^r$ and $n = p^s$ for nonnegative integers r and s. Therefore, $p(x) = x^{p^r} - c^{p^r}$ in $K[x]$. If $s > 0$, then $k = c^{p^e} = (c^{p^r})^{p^s} \in K^{p^s} \subseteq K^p$, which is contrary to the assumption $k \notin K^p$. Thus, $s = 0$ and so $r = e$. Hence, $p(x) = f(x)$, i.e., $f(x)$ is irreducible. ■

Definition 17.1.24 *Let F/K be a field extension of characteristic $p > 0$. Let $c \in F$ be a root of the irreducible polynomial $p(x)$ in $K[x]$. If the degree of separability n_0 of $p(x)$ equals 1, then c is said to be **purely inseparable** over K. If every element of F is purely inseparable over K, then F/K is called a **purely inseparable extension**.*

In Theorem 17.1.15, let c be a root of $p(x)$. Then c^{p^e} is a root of $s(y)$. We have $K(c) \supseteq K(c^{p^e}) \supseteq K$ and c is a root of the polynomial $x^{p^e} - c^{p^e}$ over $K(c^{p^e})$. It follows that $x^{p^e} - c^{p^e}$ is irreducible over $K(c^{p^e})$, $K(c)/K(c^{p^e})$ is purely inseparable, and $K(c^{p^e})/K$ is separable.

Theorem 17.1.25 *Let F/K be a field extension of characteristic $p > 0$ and c be an element of F. Then c is purely inseparable over K if and only if $c^{p^m} \in K$ for some nonnegative integer m.*

Proof. Let c be purely inseparable over K. Then the degree of separability n_0 of the minimal polynomial $p(x)$ of c equals 1. Thus, $p(x) = x^{p^e} + k$ in $K[x]$, where e is the exponent of inseparability of $p(x)$ over K. Therefore, $c^{p^e} + k = 0$ or $c^{p^e} = -k \in K$. Hence, we can take $m = e$. Conversely, suppose $c^{p^m} \in K$. Let e be the smallest nonnegative integer such that $c^{p^e} \in K$. Then c is a root of the polynomial $x^{p^e} - k$ over K, where $k = c^{p^e}$. If $x^{p^e} - k$ is not irreducible over K, then $e > 0$ and $k = k'^p$ for some $k' \in K$ by Theorem 17.1.23. In this case, $x^{p^e} - k = (x^{p^{e-1}} - k')^p$. Thus, $(c^{p^{e-1}} - k')^p = 0$ and since a field has no nonzero nilpotent elements, $c^{p^{e-1}} - k' = 0$ or $c^{p^{e-1}} = k' \in K$. However, this contradicts the minimality of e. Thus, $x^{p^e} - k$ is irreducible over K. Clearly the degree of separability of $x^{p^e} - k$ is 1. Therefore, c is purely inseparable over K. ■

Corollary 17.1.26 *Let F/K be a field extension of characteristic $p > 0$ and $c \in F$.*

(i) If c is algebraic over K, then c is purely inseparable over K if and only if the minimal polynomial of c over K is $x^{p^e} - c^{p^e}$, where e is the smallest nonnegative integer such that $c^{p^e} \in K$.

(ii) If c is purely inseparable over K, then $[K(c):K] = p^e$ for some nonnegative integer e.

(iii) If c is purely inseparable and separable algebraic over K, then $c \in K$.

[1] The remainder of this section may be skipped without any discontinuity. The only place this material is needed is in Example 24.2.8.

Proof. The proof of (i) follows from Theorem 17.1.25. Statement (ii) is an immediate consequence of statement (i). For the proof of statement (iii), we see that since c is purely inseparable over K the minimal polynomial of c over K has the form $x^{p^e} - k$. Since c is separable algebraic over K, the exponent of inseparability of $x^{p^e} - k$ is 0, i.e., $e = 0$. Thus, $x - k$ is the minimal polynomial of c over K, whence $c = k \in K$. ■

Corollary 17.1.27 *Let F/K be a field extension of characteristic $p > 0$.*

(i) If $F = K(M)$ for some subset M of F such that every element of M is purely inseparable over K, then F/K is a purely inseparable extension.

(ii) Let L be an intermediate field of F/K. Then F/K is purely inseparable if and only if F/L and L/K are purely inseparable.

(iii) The set of all elements of F which are purely inseparable over K is an intermediate field of F/K.

Proof. (i) Let c be an element of F. Then there exists a finite subset $\{m_1, m_2, \ldots, m_s\}$ of M such that

$$c = \sum_{i_1,\ldots,i_s} k_{i_1\ldots i_s} m_1^{i_1} \cdots m_s^{i_s},$$

where here we are using the fact that $F = K[M]$ since F/K is necessarily an algebraic extension. Let $e = \max\{e_1, \ldots, e_s\}$, where e_i is a nonnegative integer such that $m_i^{p^{e_i}} \in K$, $i = 1, \ldots, s$. Then

$$c^{p^e} = \sum_{i_1,\ldots,i_s} k_{i_1\ldots i_s}^{p^e} (m_1^{p^e})^{i_1} \cdots (m_s^{p^e})^{i_s} \in K.$$

Hence, c is purely inseparable over K.

(ii) Suppose that F/K is purely inseparable. Let $c \in F$. Then there exists a nonnegative integer e such that $c^{p^e} \in K$ and so $c^{p^e} \in L$. Thus, F/L is purely inseparable. L/K is purely inseparable since every element of L is an element of F. Conversely, suppose F/L and L/K are purely inseparable. Let $c \in F$. Then there exists a nonnegative integer m such that $c^{p^m} \in L$. Since L/K is purely inseparable, there exists a nonnegative integer n such that $(c^{p^m})^{p^n} \in K$. Therefore, $c^{p^{m+n}} \in K$ so that c is purely inseparable over K.

(iii) Let J denote the set of all elements of F which are purely inseparable over K. Then $K \subseteq J$ and so $J \neq \phi$. Let $c, d \in J$. Then $c^{p^e} \in K$ and $d^{p^f} \in K$ for some nonnegative integers e and f. Let $n = \max\{e, f\}$. Then $(c-d)^{p^n} = c^{p^n} - d^{p^n} \in K$. Hence, $c - d \in J$. If $d \neq 0$, then $(cd^{-1})^{p^n} = c^{p^n}(d^{p^n})^{-1} \in K$. Thus, $cd^{-1} \in J$. Hence, J is an intermediate field of F/K.

■

Theorem 17.1.25 and Corollary 17.1.27(i) make it quite easy to construct examples of purely inseparable field extensions.

Example 17.1.28 *Let J be any field of characteristic $p > 0$; e.g., $J = \mathbb{Z}_p$. Let $F = J(x, y, z)$, where x, y, z are algebraically independent over J. Set $K = J(x^p, y^{p^2}, z^{p^3})$. Then F/K is purely inseparable since x, y, z are purely inseparable over K. It can be shown that $[F : K] = p^6$ since x, y, z are algebraically independent over J. Since x^p, y^{p^2}, $z^{p^3} \in K$, we have $F^{p^3} \subseteq K$.*

For any field F of prime characteristic p, F/F^{p^e} is a purely inseparable field extension for any nonnegative integer e.

The following example is essentially the same as that in Example 17.1.28.

Example 17.1.29 *Let J be any field of characteristic $p > 0$. Let $K = J(x, y, z)$, where x, y, z are algebraically independent over J. Let $F = J(a, b, c)$, where a is a root of the polynomial $t^p - x$ over K, b is a root of the polynomial $t^{p^2} - y$ over $K(a)$, and c is a root of the polynomial $t^{p^3} - z$ over $K(a, b)$. Then F/K is purely inseparable, $[F : K] = p^6$, and $F^{p^3} \subseteq K$. One often writes $a = x^{p^{-1}}$, $b = y^{p^{-2}}$, and $c = z^{p^{-3}}$.*

Example 17.1.30 *Let J be any field of characteristic $p > 0$. Let $K = J(t)$, where t is transcendental over J. Let $F = K(t^{p^{-1}}, t^{p^{-2}}, t^{p^{-3}}, \ldots)$. Then F/K is purely inseparable by Corollary 17.1.27. Since $[K(t^{p^{-1}}, t^{p^{-2}}, t^{p^{-3}}, \ldots, t^{p^{-n}}) : K(t^{p^{-1}}, t^{p^{-2}}, t^{p^{-3}}, \ldots, t^{p^{-n+1}})] = p$ for all positive integers n, $[F : K] = \infty$. There does not exist a positive integer e such that $F^{p^e} \subseteq K$.*

Example 17.1.31 *Let J be any field of characteristic $p > 0$. Let $K = J(x_1, x_2, x_3, \ldots)$, where $x_1, x_2, x_3, \ldots$ are algebraically independent over J. Let $F_0 = K(x_1^{p^{-1}}, x_2^{p^{-2}}, x_3^{p^{-3}}, \ldots)$. Then F_0/K is purely inseparable and $[F_0 : K] = \infty$. Let $F_1 = K(x_1^{p^{-2}}, x_2^{p^{-2}}, x_3^{p^{-2}}, \ldots)$. Then F_1/K is purely inseparable, $[F_1 : K] = \infty$, and $F_1^{p^2} \subseteq K$.*

We now turn our attention to separable extensions.

Theorem 17.1.32 *Let F/K be a field extension of characteristic $p > 0$. If F/K is separable algebraic, then $F = K(F^p)$. If $[F : K] < \infty$ and $F = K(F^p)$, then F/K is separable algebraic.*

Proof. Suppose F/K is separable algebraic. Now every element of F is purely inseparable over F^p and thus purely inseparable over $K(F^p)$. Every element c of F is separable algebraic over K and thus separable algebraic over $K(F^p)$. Thus, every element c of F is in $K(F^p)$ by Corollary 17.1.26(iii). Hence, $F \subseteq K(F^p)$, so that $F = K(F^p)$. Conversely, suppose $[F : K] < \infty$ and $F = K(F^p)$. Let a be any element of F. Since $[F : K] < \infty$, a is algebraic over K. If a is not separable over K, then the minimal polynomial of a over K has the form

$$(x^p)^n + \cdots + k_1 x^p + k_0.$$

Therefore, $0 = a^{np} + \cdots + k_1 a^p + k_0 \cdot 1$ with not all the $k_i = 0$. Hence, $1, a^p, \ldots, a^{np}$ are linearly dependent over K. By Theorem 16.1.14, $1, a, a^2, \ldots, a^n, \ldots, a^{np-1}$ are linearly independent over K, whence $1, a, a^2, \ldots, a^n$ are linearly independent over K.

We now show that this is impossible by showing that whenever n elements $b_1, \ldots, b_n$ of F are linearly independent over K, then the elements $b_1^p, \ldots, b_n^p$ are linearly independent over K. We can assume that $b_1, \ldots, b_n$ is a basis of F/K since any linearly independent set over K can be extended to a basis of F/K, in particular, the linearly independent set $\{1, a, \ldots, a^n\}$. By Exercise 7 (page 256), the mapping $\alpha : F \to F^p$ defined by $\alpha(c) = c^p$ for $c \in F$ is an isomorphism, which maps K onto K^p. Thus, since $b_1, \ldots, b_n$ is a basis of F/K, $b_1^p, \ldots, b_n^p$ is a basis of F^p/K^p. Hence, $b_1^p, \ldots, b_n^p$ spans F^p over K^p. Consequently, $b_1^p, \ldots, b_n^p$ spans $K(F^p)$ over K; i.e., F over K. Since F has dimension n over K and the n elements $b_1^p, \ldots, b_n^p$ span F over K, the elements $b_1^p, \ldots, b_n^p$ must be a basis for F over K. ■

The field extension F/K of Example 17.1.30 shows that the finiteness condition $[F : K] < \infty$ cannot be dropped in the above theorem. We have $F = K(F^p)$, F/K is not separable algebraic, in fact, F/K is purely inseparable.

Corollary 17.1.33 *Let F/K be a field extension of characteristic $p > 0$.*

(i) Let a be an element of F. Then $K(a) = K(a^p)$ if and only if $K(a)/K$ is separable algebraic.

(ii) Let $a_1, a_2, \ldots, a_n$ be elements of F. Then $K(a_1, \ldots, a_n)/K$ is separable algebraic if and only if a_1 is separable algebraic over K and a_i is separable algebraic over $K(a_1, \ldots, a_{i-1})$, $i = 2, 3, \ldots, n$.

Proof. (i) If $K(a) = K(a^p)$, then a cannot be transcendental over K and so a must be algebraic over K. By Theorem 17.1.32, $K(a) = K(K(a)^p)$ if and only if $K(a)/K$ is separable algebraic. We thus have the desired result since $K(K(a)^p) = K(a^p)$.

(ii) Suppose $K(a_1, \ldots, a_n)/K$ is separable algebraic. Then $a_1, \ldots, a_n$ are separable algebraic over K. By the discussion following Definition 17.1.2, a_i is clearly separable algebraic over $K(a_1, \ldots, a_{i-1})$, $i = 2, 3, \ldots, n$. Conversely, suppose a_1 is separable algebraic over K and a_i is separable algebraic over $K(a_1, \ldots, a_{i-1})$, $i = 2, 3, \ldots, n$. Then $K(a_1) = K(a_1^p)$, $\ldots$, $K(a_1, \ldots, a_{i-1})(a_i) = K(a_1, \ldots, a_{i-1})(a_i^p)$, $i = 2, 3, \ldots, n$. Thus, $K(a_1, \ldots, a_n) = K(a_1^p, \ldots, a_n^p) = K([K(a_1, \ldots, a_{i-1})]^p)$. The conclusion now holds from Theorem 17.1.32. ■

Corollary 17.1.34 *Let F/K be a field extension of characteristic $p > 0$.*

(i) If $F = K(M)$ for some subset M of F such that every element of M is separable algebraic over K, then F/K is separable algebraic.

(ii) Let L be an intermediate field of F/K. Then F/K is separable algebraic if and only if F/L and L/K are separable algebraic.

(iii) The set of all elements of F which are separable algebraic over K is an intermediate field of F/K.

Proof. (i) Let $a \in F$. There exists a finite subset $\{m_1, \ldots, m_s\}$ of M such that $a \in K(m_1, \ldots, m_s)$. Since each m_i is separable algebraic over K, we have by Corollary 17.1.33(ii) that $K(m_1, \ldots, m_s)/K$ is separable algebraic. Hence, a and thus F/K is separable algebraic.

(ii) Suppose F/K is separable algebraic. Then F/L is separable algebraic by the discussion following Definition 17.1.2. L/K is separable algebraic since every element of L is an element of F. Suppose F/L and L/K are separable algebraic. Let $a \in F$. Let $c_0, c_1, \ldots, c_n \in L$ be the coefficients of the minimal polynomial $p(x)$ of a over L. Since a is separable algebraic over L, a is separable algebraic over $K(c_0, c_1, \ldots, c_n)$. ($p(x)$ is also the minimal polynomial of a over $K(c_0, c_1, \ldots, c_n)$.) Since $c_0, c_1, \ldots, c_n \in L$ and L/K is separable algebraic, $K(c_0, c_1, \ldots, c_n)/K$ is separable algebraic by Corollary 17.1.33(ii). Thus, a and so F is separable algebraic over K.

(iii) Let S denote the set of elements of F which are separable algebraic over K. Then $S \supseteq K$. Let $a, b \in S$. Then by Corollary 17.1.33(ii), $K(a,b)/K$ is separable algebraic. Since $a - b \in K(a,b)$ and (for $b \neq 0$) $ab^{-1} \in K(a,b)$, $a-b$, and ab^{-1} $(b \neq 0)$ are separable algebraic over K and thus are members of S. Hence, S is a field. ■

Definition 17.1.35 *Let F/K be an algebraic field extension of characteristic $p > 0$. Then the intermediate field of F/K consisting of all elements of F which are separable algebraic over K is called the* ***separable closure*** *of K in F or the* ***maximal separable intermediate*** *field of F/K. We denote this field by K_s.*

Theorem 17.1.36 *Let F/K be an algebraic field extension of characteristic $p > 0$. Then F/K_s is purely inseparable, where K_s is the separable closure of F/K.*

Proof. If $F = K_s$ the theorem is immediate. Suppose $F \supset K_s$. Let $a \in F, a \notin K_s$. Let

$$p(x) = k_0 + k_1 x^{p^e} + \cdots + (x^{p^e})^{n_0}$$

be the minimal polynomial of F/K_s, where e is the exponent of inseparability and n_0 is the reduced degree of $p(x)$ over K_s. Now by Theorem 17.1.15, $k_0 + k_1 y + \cdots + y^{n_0}$ is the minimal polynomial of a^{p^e} over K_s and this polynomial is separable over K_s. Hence, a^{p^e} is separable over K_s. Thus, $K_s(a^{p^e})/K_s$ is separable algebraic and so $K_s(a^{p^e})/K$ is separable algebraic. By the definition of K_s, we have $a^{p^e} \in K_s$. Therefore, a is purely inseparable over K_S.
■

We can think of field theory as being separated into two parts, namely, that in which the fields are of characteristic 0 and that in which the fields are of prime characteristic p. It can be shown that for any field extension F/K, there exists a subset X of F which is algebraically independent over K and which also has the property that $F/K(X)$ is algebraic. The above theorem shows that the study of algebraic field extensions of characteristic $p > 0$ can be separated into two parts, the separable part and the purely inseparable part. Separable algebraic field extensions of characteristic $p > 0$ often act entirely similar to field extensions of characteristic 0. Purely inseparable field extensions have their own distinctive behavior.

Definition 17.1.37 *Let F/K be an algebraic field extension of characteristic $p > 0$. Then the degree $[K_s : K]$ is called the* ***degree of separability*** *of F/K and is denoted by $[F : K]_s$. The degree $[F : K_s]$ is called the* ***degree of inseparability*** *of F/K and is denoted by $[F : K]_i$.*

Theorem 17.1.38 *Let K be a field of characteristic $p > 0$ and $p(x)$ an irreducible polynomial in $K[x]$. Let $K(a)$ be an extension of K obtained by adjoining a root a of $p(x)$ to K. Then*

$$[K(a) : K]_s = n_0,$$

$$[K(a) : K]_i = p^e,$$

where n_0 is the reduced degree of $p(x)$ over K and p^e is the degree of inseparability of $p(x)$ over K.

Proof. Let $b \in K(a)$. Then $b = \sum_{i=0}^{n-1} k_i a^i$, where n is the degree of $p(x)$ over K and each $k_i \in K$. Therefore,

$$b^{p^e} = \sum_{i=0}^{n-1} k_i^{p^e} (a^{p^e})^i \in K(a^{p^e}).$$

Thus, b is purely inseparable over $K(a^{p^e})$. Hence, $K(a)/K(a^{p^e})$ is purely inseparable. By the definition of the degree of inseparability of $p(x)$ over K, $K(a^{p^e})/K$ is separable algebraic. Now $K_s \supseteq K(a^{p^e})$. Let $b \in K_s$. We have just seen that b is purely inseparable over $K(a^{p^e})$. But b is also separable algebraic over $K(a^{p^e})$. Therefore, $b \in K(a^{p^e})$ so that $K_s = K(a^{p^e})$. By Theorem 17.1.15, the minimal polynomial of a^{p^e} over K is of degree n_0 and so $[K(a) : K]_s = [K(a^{p^e}) : K] = n_0$. Thus, $n_0 p^e = [K(a) : K] = [K(a) : K(a^{p^e})][K(a^{p^e}) : K] = [K(a) : K(a^{p^e})]n_0$. Consequently, $p^e = [K(a) : K(a^{p^e})] = [K(a) : K]_i$. ■

Example 17.1.39 *Let K denote the field $\mathbb{Z}_p(u,v)$, where u and v are algebraically independent over $\mathbb{Z}_p$. Let a be a root of the polynomial $x^{2p} + vx^p + u$ over K. By use of Worked-Out Exercise 6 (page 236), one can deduce that $x^{2p} + vx^p + u$ is irreducible over K. Let F be the field $K(a)$. We ask the reader to verify the following properties of the field extension F/K. $K_s = K(a^p)$, $[F : K]_i = p$, and $[F : K]_s = 2$. Also, the extension F/K has no elements which are purely inseparable over K (except those elements which are already in K). Thus, if J is the intermediate field of F/K consisting of all the elements of F purely inseparable over K, then $J = K$. Hence, F/J is not separable algebraic.*

Worked-Out Exercises

♢ **Exercise 1**: Determine if the following polynomials are separable or inseparable over the given fields.

(a) $x^2 - 6x + 9$ over $\mathbb{Q}$;

(b) $x^4 + x^2 + [1]$ over $\mathbb{Z}_2$.

Solution: (a) $x^2 - 6x + 9 = (x-3)^2$ over $\mathbb{Q}$. Now $x - 3$ is irreducible over $\mathbb{Q}$. Since $x - 3$ is separable over $\mathbb{Q}$, $x^2 - 6x + 9$ is separable over $\mathbb{Q}$.

(b) $x^4 + x^2 + [1] = (x^2 + x + [1])^2$ over $\mathbb{Z}_2$. Now $x^2 + x + [1]$ has no roots in $\mathbb{Z}_2$. Hence, $x^2 + x + [1]$ is irreducible over $\mathbb{Z}_2$. Now $D_x(x^2 + x + [1]) = [2]x + [1] = [1] \neq [0]$. Thus, $x^2 + x + [1]$ and so $x^4 + x^2 + [1]$ is separable over $\mathbb{Z}_2$.

♢ **Exercise 2**: Prove that the following polynomials are irreducible over $\mathbb{Z}_3(t)$, where t is transcendental over $\mathbb{Z}_3$. Find the exponent of inseparability and the degree of separability of the polynomials over $\mathbb{Z}_3(t)$.

(a) $p(x) = x^{36} + tx^{18} + t$.

(b) $q(x) = x^{24} + tx^{18} + t$.

(c) $r(x) = x^{20} + tx^{18} + t$.

(d) $s(x) = x^9 + t$.

Solution: Since $t|t$, $t|0$, $t \not| 1$, $t^2 \not| t$, the polynomials $p(x)$, $q(x)$, $r(x)$, $s(x)$ are irreducible over $\mathbb{Z}_3(t)$.

(a) $p(x) = x^{4 \cdot 3^2} + tx^{2 \cdot 3^2} + t$ and so the exponent of inseparability $e = 2$ and the degree of separability $n_0 = 4$.

(b) $q(x) = x^{8 \cdot 3} + tx^{6 \cdot 3} + t$ and so the exponent of inseparability $e = 1$ and the degree of separability $n_0 = 8$.

(c) Since $3 \not| 20$, $e = 0$ and $n_0 = 20$.

(d) Here $e = 2$ and $n_0 = 1$.

♢ **Exercise 3**: Let $f(x)$ and $g(x)$ be polynomials over the field K.

(a) Does $f(c) = g(c)$ for all $c \in K$ imply that $f(x) = g(x)$?

(b) Does $f(c) = 0$ for all $c \in K$ imply that $f(x) = 0$?

Solution: (a) Let $f(x) = [3]x^5 - [4]x^2 \in \mathbb{Z}_5[x]$ and $g(x) = x^2 + [3]x \in \mathbb{Z}_5[x]$. Now $f([0]) = [0] = g([0])$, $f([1]) = [4] = g([1])$, $f([2]) = [0] = g([2])$, $f([3]) = [3] = g([3])$, $f([4]) = [3] = g([4])$. Hence, $f(c) = g(c)$ for all $c \in \mathbb{Z}_5$. However, $f(x) \neq g(x)$.

(b) Let $f(x) = x^2 + x \in \mathbb{Z}_2[x]$. Then $f(c) = 0$ for all $c \in \mathbb{Z}_2$, but $f(x) \neq 0$.

Exercise 4: Let $K = P(x, y, z)$ and $F = K(z^{p^{-2}}, z^{p^{-2}}x^{p^{-1}} + y^{p^{-1}})$, where P is a perfect field of characteristic $p > 0$ and x, y, z are algebraically independent indeterminates over P. Prove that $K^{p^{-1}} \cap F = K(z^{p^{-1}})$, where $K^{p^{-1}} = \{k^{p^{-1}} \mid k \in K\}$.

Solution: Clearly $F \supset K^{p^{-1}} \cap F \supseteq K(z^{p^{-1}})$. Now $[F : K] = p^3$. Suppose that $K^{p^{-1}} \cap F \supset K(z^{p^{-1}})$. Then $F = (K^{p^{-1}} \cap F)(z^{p^{-2}})$ since $z^{p^{-2}} \notin K^{p^{-1}} \cap F$ and $[K^{p^{-1}} \cap F : K]$ must be p^2. Thus, $[F : K^{p^{-1}} \cap F] = p$. Since $[K^{p^{-1}}(F) : K^{p^{-1}}] = p$, any basis of $F/(K^{p^{-1}} \cap F)$ remains a basis of $K^{p^{-1}}(F)/K^{p^{-1}}$. Now $Z = \{1, z^{p^{-2}}, \ldots, (z^{p^{-2}})^{p-1}\}$ is a basis of $F/(K^{p^{-1}} \cap F)$. Also,

$$z^{p^{-2}}x^{p^{-1}} + y^{p^{-1}} = \sum_{i=0}^{p-1} k_i (z^{p^{-2}})^i,$$

where $k_i \in K^{p^{-1}} \cap F$, $i = 0, 1, \ldots, p-1$. Since Z remains linearly independent over $K^{p^{-1}}$, $y^{p^{-1}} = k_0 \in K^{p^{-1}} \cap F$ and $x^{p^{-1}} = k_1 \in K^{p^{-1}} \cap F$. Therefore, $x^{p^{-1}}, y^{p^{-1}} \in F$. Thus, $[F : K] = p^4$, a contradiction. Hence, $K^{p^{-1}} \cap F = K(z^{p^{-1}})$.

Exercises

1. Let $f(x) \in K[x]$, a polynomial ring over a field K and $c \in F$, where F is an extension field of K. Prove that $(x-c)^2 | f(x)$ if and only $(x-c)|f(x)$ and $(x-c)|f'(x)$.
2. Let $f(x) \in K[x]$, a polynomial ring over a field K. Use Exercise 1 to prove that $f(x)$ has no repeated roots in any extension field of K if and only if $f(x)$ and $f'(x)$ are relatively prime.
3. Let $f(x) = x^n - x \in K[x]$, a polynomial ring over a field K. Suppose that $n \geq 2$ and that either K has characteristic 0 or a prime p such that p does not divide $n-1$. Prove that $f(x)$ has no repeated roots in any extension field F of K.
4. Let $f(x) = x^p - k \in K[x]$, a polynomial ring over a field K of characteristic $p > 0$. Prove that either $f(x)$ is irreducible over K or that $f(x)$ is a power of a linear polynomial in $K[x]$.
5. Determine if the following polynomials are separable or inseparable over the given field.
   (i) $x^2 - 4x + 4$ over $\mathbb{Q}$.
   (ii) $x^5 + tx + t$ over $\mathbb{Z}_5(t)$, where t is transcendental over $\mathbb{Z}_5$.
6. Prove that the following polynomials are irreducible over $\mathbb{Z}_5(u)$, where u is transcendental over $\mathbb{Z}_5$. Find the exponent of inseparability and the degree of separability of the polynomials over $\mathbb{Z}_5(u)$.
   (i) $p(x) = x^{250} + ux^{125} + u$.
   (ii) $g(x) = x^{128} + ux^{125} + u$.
   (iii) $s(x) = x^{125} + u$.
7. Let F be a field of characteristic $p > 0$. Prove that for any nonnegative integer e, F^{p^e} is a subfield of F. Prove also that the mapping $\alpha : F \to F^{p^e}$ defined by $\alpha(a) = a^{p^e}$ is an isomorphism.
8. Prove that a root of the polynomials in Examples 17.1.16 and 17.1.39 is neither purely inseparable nor separable algebraic over $K(t)$ and K, respectively.
9. Let $K(a)/K$ be a field extension of characteristic $p > 0$. Prove that $(K(a))^p = K^p(a^p)$.
10. Let F/K be a finite field extension of characteristic $p > 0$. If $[F : K]$ is not divisible by p, prove that F/K is separable.
11. Let F/K be an algebraic field extension and S be an intermediate field of F/K such that F/S is purely inseparable and S/K is separable algebraic. Prove that $S = K_s$.
12. Let P be a perfect field of characteristic $p > 0$. Let $P(a)/P$ be an algebraic field extension. Prove that $P(a)/P$ is separable and that $P(a)$ is perfect.
13. Let K be any field of characteristic $p > 0$. Prove that $\mathbb{Z}_p$ is the smallest subfield of K which is perfect and $\cap_{i=0}^{\infty} K^{p^i}$ is the largest subfield of K which is perfect.
14. Verify the properties of the field extension F/K of Example 17.1.39.
15. Answer the following statements, true or false. If the statement is true, prove it. If it is false, give a counterexample.
    (i) Let F be a field of characteristic $p > 0$. Since $F \simeq F^p$ and $F^p \subseteq F$, it follows that $F^p = F$.
    (ii) Let F/K be a field extension of characteristic $p > 0$. Let $c \in F \backslash K$. Then it is impossible for c to be both separable and purely inseparable over K.
    (iii) Let F/K be a field extension of characteristic $p > 0$. Let $c \in F$. Then it is impossible for c to be both separable and inseparable over K.

Chapter 18

Finite Fields

The theory of finite fields has come to the fore in the last 60 years due to newfound applications. The applications of finite fields are in coding theory, combinatorics, switching circuits, statistics via finite geometries, and certain areas of computer science.

18.1 Finite Fields

A **finite field** (or **Galois field**) is a field with a finite number of elements. If F is a finite field, then F has prime characteristic p and contains a subfield isomorphic to $\mathbb{Z}_p$. Since F has only a finite number of elements, $[F : \mathbb{Z}_p] < \infty$.

We denote a finite field of n elements by GF(n). We will show in the next result that n must be a power of p. The result is due to E.H. Moore (1862–1932). The United States is indebted to Moore for its beginnings in abstract algebra and for its initial international recognition in research.

Theorem 18.1.1 *If F is a finite field of characteristic p and $n = [F : \mathbb{Z}_p]$, then F contains p^n elements.*

Proof. Since $[F : \mathbb{Z}_p] = n$, $F/\mathbb{Z}_p$ has a basis of n elements, say, $b_1, b_2, \ldots, b_n$. Every element a of F is a linear combination of b_1, b_2, $\ldots$, b_n, i.e., $a = a_1b_1 + a_2b_2 + \cdots + a_nb_n$, where $a_i \in \mathbb{Z}_p$, $i = 1, 2, \ldots, n$. Now $\mathbb{Z}_p$ has p elements. Hence, F has at most p^n elements. Since $\{b_1, b_2, \ldots, b_n\}$ is linearly independent over $\mathbb{Z}_p$, $a_1b_1 + a_2b_2 + \cdots + a_nb_n$ is distinct for every choice of $a_1, a_2, \ldots, a_n$. Thus, F has exactly p^n elements. ■

Theorem 18.1.2 *Every element of a finite field F of characteristic p and of p^n elements is a root of the polynomial $x^{p^n} - x \in \mathbb{Z}_p[x]$. Moreover, F is a splitting field of $x^{p^n} - x$ over $\mathbb{Z}_p$.*

Proof. First note that $(F\backslash\{0\}, \cdot)$ is a commutative group of order $p^n - 1$. Thus, for all $a \in F\backslash\{0\}$, $a^{p^n-1} = 1$, whence $a^{p^n} = a$. Clearly $0^{p^n} = 0$. Since F contains all the roots of $x^{p^n} - x$, F contains a splitting field S of $x^{p^n} - x$ over $\mathbb{Z}_p$. However, F is exactly the set of all the roots of $x^{p^n} - x$ and so $F = S$.
■

In the following result, we once again use a positive integer and the concept of an isomorphism to completely characterize an algebraic structure.

Corollary 18.1.3 *Any two finite fields of p^n elements are isomorphic, where p is a prime and n is a positive integer.*

Proof. If F and F' are finite fields with p^n elements, then they are splitting fields of the polynomial $x^{p^n} - x$ over $\mathbb{Z}_p$. Hence, $F \simeq F'$.
■

The next theorem can be used to show that there exists an irreducible polynomial of arbitrary degree n over $\mathbb{Z}_p$. (See Exercise 8, page 260.) Even though its proof is not constructive in nature, it is informative for certain applications. Exercises 5 and 6 can be used to actually count the irreducible polynomials of a given degree. There is an algorithm which can be used to test the irreducibility of a polynomial over a finite field—namely, Berlekamp's algorithm. This algorithm is discussed in Isaacs.

Theorem 18.1.4 *For any prime p, there exists a field extension $F/\mathbb{Z}_p$ of arbitrary finite degree n.*

Proof. Let S be the splitting field of the polynomial $f(x) = x^{p^n} - x$ over $\mathbb{Z}_p$. Let $a \in S$ be a root of $f(x)$ of multiplicity m. Then

$$f(x) = (x-a)^m g(x),$$

where a is not a root of $g(x)$. Now

$$-1 = f'(x) = (x-a)^{m-1}[mg(x) + (x-a)g'(x)].$$

This implies that $(x-a)^{m-1}$ divides -1, whence $m-1=0$. Thus, every root of $f(x)$ in S has multiplicity 1. Hence, $f(x)$ has p^n distinct roots in S. Let F denote the subset of S, which consists of all roots of $f(x)$. Let a, $b \in F$. Then $(a-b)^{p^n} = a^{p^n} - b^{p^n} = a - b$. Therefore, $a - b \in F$. For $b \neq 0$,

$$(ab^{-1})^{p^n} = a^{p^n}(b^{p^n})^{-1} = ab^{-1}$$

and so $ab^{-1} \in F$. Thus, F is a subfield of S. Since F contains all the roots of $f(x)$ and S is generated by the roots of $f(x)$ over $\mathbb{Z}_p$, $F = S$. By Exercise 6 (page 260), $[F : \mathbb{Z}_p] = n$. ■

Theorem 18.1.5 *Let F be a field and G be a finite subgroup of the multiplicative group $F^* = F\backslash\{0\}$. Then G is cyclic.*

Proof. Since G is a finite Abelian group, G is a direct product of cyclic subgroups $C_1, C_2, \ldots, C_k$, where $|C_i| = n_i$, $n_1 > 1$, and $n_i | n_{i+1}$, $1 \leq i < k$. From this it follows that $g^{n_k} = 1$ for all $g \in G$. Thus, every element of G is a root of $x^{n_k} - 1 \in F[x]$. Since $x^{n_k} - 1$ has at most n_k distinct roots in F, $|G| \leq n_k$. Now C_k is a subgroup of G and $|C_k| = n_k$. Hence, $G = C_k$ and so G is cyclic.
■

The following corollary is an immediate consequence of Theorem 18.1.5.

Corollary 18.1.6 *The multiplicative group of a finite field is cyclic.* ■

Theorem 18.1.7 *Let F be a finite field and $F(a, b)/F$ a field extension with a, b algebraic over F. Then there exists $c \in F(a, b)$ such that $F(a, b) = F(c)$, i.e., $F(a, b)$ is a simple extension.*

Proof. Since $F(a, b)/F$ is algebraic, $[F(a, b) : F] < \infty$. Thus, $F(a, b)$ is a finite field since F is a finite field. Since $F(a, b)\backslash\{0\}$ is a cyclic group with some generator, say, c by Theorem 18.1.5, it follows that $F(a, b) = F(c)$.
■

Corollary 18.1.8 *Every finite extension of a finite field is simple.* ■

Worked-Out Exercises

♢ **Exercise 1**: Prove that $x^3 + x + [1]$ is irreducible in $\mathbb{Z}_2[x]$. Write out the addition and multiplication tables for the field

$$\mathbb{Z}_2[x]/\langle x^3 + x + [1]\rangle.$$

Find a splitting field S_1 for $x^3 + x + [1]$ over $\mathbb{Z}_2$. Find a basis for $S_1/\mathbb{Z}_2$ and $[S_1 : \mathbb{Z}_2]$.

Solution: $x^3 + x + [1]$ is irreducible over $\mathbb{Z}_2$ if and only if $\mathbb{Z}_2$ contains no root of $x^3 + x + [1]$. Since $[0]^3 + [0] + [1] \neq [0]$ and $[1]^3 + [1] + [1] \neq [0]$ in $\mathbb{Z}_2$, $\mathbb{Z}_2$ contains no roots of $x^3 + x + [1]$ over $\mathbb{Z}_2$. Hence, $x^3 + x + [1]$ is irreducible over $\mathbb{Z}_2$. By Theorem 16.1.11,

$$\mathbb{Z}_2[x]/\langle x^3 + x + [1]\rangle = \mathbb{Z}_2(\lambda),$$

where λ denotes the coset $x + \langle x^3 + x + [1]\rangle$. By Theorem 16.1.14,

$$\mathbb{Z}_2(\lambda) = \{[0], [1], \lambda, \lambda^2, [1] + \lambda, [1] + \lambda^2, \lambda + \lambda^2, [1] + \lambda + \lambda^2\}.$$

Now

$$x^3 + x + [1] = (x + \lambda)(x^2 + \lambda x + [1] + \lambda^2)$$

and λ^2 and $\lambda + \lambda^2$ are the roots of $x^2 + \lambda x + [1] + \lambda^2$. Since λ^2, $\lambda + \lambda^2 \in \mathbb{Z}_2(\lambda)$, $\mathbb{Z}_2(\lambda)$ is a splitting field of $x^3 + x + [1]$ over $\mathbb{Z}_2$. Let $S_1 = \mathbb{Z}_2(\lambda)$. Then $\{[1], \lambda, \lambda^2\}$ is a basis for $S_1/\mathbb{Z}_2$ and $[S_1 : \mathbb{Z}_2] = 3$. Let α denote

$[1]+\lambda+\lambda^2$. The addition table for $\mathbb{Z}_2(\lambda)$ is given below.

+	[0]	[1]	λ	λ^2	[1]+λ	[1]+λ^2	λ+λ^2	α
[0]	[0]	[1]	λ	λ^2	[1]+λ	[1]+λ^2	λ+λ^2	α
[1]	[1]	[0]	[1]+λ	[1]+λ^2	λ	λ^2	α	λ+λ^2
λ	λ	[1]+λ	[0]	λ+λ^2	[1]	α	λ^2	[1]+λ^2
λ^2	λ^2	[1]+λ^2	λ+λ^2	[0]	α	[1]	λ	[1]+λ
[1]+λ	[1]+λ	λ	[1]	α	[0]	λ+λ^2	[1]+λ^2	λ^2
[1]+λ^2	[1]+λ^2	λ^2	α	[1]	λ+λ^2	[0]	[1]+λ	λ
λ+λ^2	λ+λ^2	α	λ^2	λ	[1]+λ^2	[1]+λ	[0]	[1]
α	α	λ+λ^2	[1]+λ^2	[1]+λ	λ^2	λ	[1]	[0]

For the multiplication table, we make a few entries, such as $([1]+\lambda)([1]+\lambda)=[1]+\lambda^2$ and $([1]+\lambda+\lambda^2)([1]+\lambda^2)=[1]+\lambda+\lambda^3+\lambda^4$. We now reduce $[1]+\lambda+\lambda^3+\lambda^4$ to the form $a+b\lambda+c\lambda^2$, where $a,b,c\in\mathbb{Z}_2$. We divide $x^4+x^3+x+[1]$ by $x^3+x+[1]$ to obtain $x^4+x^3+x+[1]=(x+[1])(x^3+x+[1])+x^2+x$. Thus, $\lambda^4+\lambda^3+\lambda+[1]=(\lambda+[1])(\lambda^3+\lambda+[1])+\lambda^2+\lambda=[0]+\lambda^2+\lambda$. Hence, $([1]+\lambda+\lambda^2)([1]+\lambda^2)=\lambda+\lambda^2$.

♢ **Exercise 2**: Prove that $x^3+x^2+[1]$ is irreducible in $\mathbb{Z}_2[x]$. Write out the addition and multiplication tables for the field

$$\mathbb{Z}_2[x]/\left\langle x^3+x^2+[1]\right\rangle.$$

Find a splitting field S_2 for $x^3+x+[1]$ over $\mathbb{Z}_2$. Find a basis for $S_2/\mathbb{Z}_2$ and $[S_2:\mathbb{Z}_2]$. Compare your results with those in Worked-Out Exercise 1.

Solution: Since $[0]^3+[0]^2+[1]\neq[0]$ and $[1]^3+[1]^2+[1]\neq[0]$ in $\mathbb{Z}_2$, $\mathbb{Z}_2$ contains no roots of $x^3+x^2+[1]$ over $\mathbb{Z}_2$. Hence, $x^3+x^2+[1]$ is irreducible over $\mathbb{Z}_2$. By Theorem 16.1.11,

$$\mathbb{Z}_2[x]/\left\langle x^3+x^2+[1]\right\rangle=\mathbb{Z}_2(\mu),$$

where μ denotes the coset $x+\left\langle x^3+x^2+[1]\right\rangle$. By Theorem 16.1.14,

$$\mathbb{Z}_2(\mu)=\{[0],[1],\mu,\mu^2,[1]+\mu,[1]+\mu^2,\mu+\mu^2,[1]+\mu+\mu^2\}.$$

Now $x^3+x^2+[1]=(x+\mu)(x^2+([1]+\mu)x+\mu+\mu^2)$ and μ^2 and $[1]+\mu+\mu^2$ are the roots of $x^2+([1]+\mu)x+\mu+\mu^2$. Since μ^2, $[1]+\mu+\mu^2\in\mathbb{Z}_2(\mu)$, $\mathbb{Z}_2(\mu)$ is a splitting field of $x^3+x^2+[1]$ over $\mathbb{Z}_2$. Let $S_2=\mathbb{Z}_2(\mu)$. Then $\{[1],\mu,\mu^2\}$ is a basis for $S_2/\mathbb{Z}_2$ and $[S_2:\mathbb{Z}_2]=3$. The addition table for $\mathbb{Z}_2(\mu)$ is determined in a manner similar to that in Exercise 1. In fact, one may obtain the addition table by substituting μ for λ in the addition table of $\mathbb{Z}_2(\lambda)$. We now consider multiplication. We note that $([1]+\mu)([1]+\mu)=[1]+\mu^2$. However, $([1]+\mu+\mu^2)([1]+\mu^2)=[1]+\mu+\mu^3+\mu^4=[1]$. Hence, we note the first algebraic difference between $\mathbb{Z}_2(\lambda)$ and $\mathbb{Z}_2(\mu)$.

♢ **Exercise 3**: Show that there exists an isomorphism f of $\mathbb{Z}_2(\lambda)$ onto $\mathbb{Z}_2(\mu)$ considered as vector spaces over $\mathbb{Z}_2$ such that f is the identity on $\mathbb{Z}_2$ and $f(\lambda)=\mu$, $f(\lambda^2)=\mu^2$, where λ and μ are as defined in Worked-Out Exercises 1 and 2, respectively.

Solution: $\{[1],\lambda,\lambda^2\}$ is a basis for $\mathbb{Z}_2(\lambda)$ over $\mathbb{Z}_2$ and $\{[1],\mu,\mu^2\}$ is a basis for $\mathbb{Z}_2(\mu)$ over $\mathbb{Z}_2$. Hence, there exists a unique linear transformation f of $\mathbb{Z}_2(\lambda)$ onto $\mathbb{Z}_2(\mu)$ such that $f([1])=[1]$, $f(\lambda)=\mu$, and $f(\lambda^2)=\mu^2$. This linear transformation is given by

$$f(a[1]+b\lambda+c\lambda^2)=a[1]+b\mu+c\mu^2,$$

where $a,b,c\in\mathbb{Z}_2$. Since $\{[1],\mu,\mu^2\}$ is linearly independent, f is one-one.

♢ **Exercise 4**: Show that $\mathbb{Z}_2(\lambda)$ and $\mathbb{Z}_2(\mu)$ are isomorphic as fields, where λ and μ are as defined in Worked-Out Exercises 1 and 2, respectively.

Solution: Since $|\mathbb{Z}_2(\lambda)|=2^3=|\mathbb{Z}_2(\mu)|$, $\mathbb{Z}_2(\lambda)$ and $\mathbb{Z}_2(\mu)$ are splitting fields of x^8-x over $\mathbb{Z}_2$ and thus are isomorphic.

♢ **Exercise 5**: Factor the polynomial x^8-x over $\mathbb{Z}_2$.

Solution: $x^8-x=x(x+[1])(x^6+x^5+x^4+x^3+x^2+x+[1])$. Now $x^2+x+[1]$ is the only irreducible quadratic polynomial over $\mathbb{Z}_2$. But $x^2+x+[1]$ does not divide $x^6+x^5+x^4+x^3+x^2+x+[1]$. We have that $x^3+x+[1]$ and $x^3+x^2+[1]$ are irreducible polynomials over $\mathbb{Z}_2$ and $x^6+x^5+x^4+x^3+x^2+x+[1]=(x^3+x+[1])(x^3+x^2+[1])$. Hence, $x^8-x=x(x+[1])(x^3+x+[1])(x^3+x^2+[1])$.

♢ **Exercise 6**: Find the roots of $x^3+x^2+[1]$ in $\mathbb{Z}_2(\lambda)$, where λ is as defined in Worked-Out Exercise 1.

Solution: [0] is a root of x, [1] is a root of $x+[1]$, and λ,λ^2, $\lambda+\lambda^2$ are roots of $x^3+x+[1]$. Hence, $[1]+\lambda$, $[1]+\lambda^2$, and $[1]+\lambda+\lambda^2$ are roots of $x^3+x^2+[1]$.

♢ **Exercise 7**: Find the roots of $x^3 + x + [1]$ in $\mathbb{Z}_2(\mu)$, where μ is as defined in Worked-Out Exercise 2.

Solution: $[0]$ is a root of x, $[1]$ is a root of $x + [1]$, and $\mu, \mu^2, [1] + \mu + \mu^2$ are roots of $x^3 + x^2 + [1]$. Hence, $[1] + \mu$, $[1] + \mu^2$, and $\mu + \mu^2$ are roots of $x^3 + x + [1]$.

♢ **Exercise 8**: Show that there exists an isomorphism g of $\mathbb{Z}_2(\lambda)$ onto $\mathbb{Z}_2([1] + \mu)$ such that $g(\lambda) = [1] + \mu$, where λ and μ are as defined in Worked-Out Exercises 1 and 2, respectively.

Solution: The result here follows immediately by Corollary 16.2.9.

♢ **Exercise 9**: Show that there does not exist an isomorphism h of $\mathbb{Z}_2(\lambda)$ onto $\mathbb{Z}_2(\mu)$ such that $h(\lambda) = \mu$, where λ and μ are as defined in Worked-Out Exercises 1 and 2, respectively.

Solution: Suppose there exists an isomorphism h of $\mathbb{Z}_2(\lambda)$ onto $\mathbb{Z}_2(\mu)$ such that $h(\lambda) = \mu$. Then $[0] = h([0]) = h(\lambda^3 + \lambda + [1]) = \mu^3 + \mu + [1]$. Also, $[0] = \mu^3 + \mu^2 + [1]$. Hence, $\mu^3 + \mu + [1] = \mu^3 + \mu^2 + [1]$. Thus, $\mu^2 = \mu$. Therefore, $\mu = [1]$, a contradiction.

Exercises

1. Let F be a finite field. A generator for $F^* = F\backslash\{0\}$ is called a **primitive element** for F. Find a primitive element for the following fields.
   (i) $\mathbb{Z}_7$.
   (ii) $\mathbb{Z}_{11}$.
   (iii) F, where $F \supseteq \mathbb{Z}_2$ and $[F : \mathbb{Z}_2] = 8$.
2. Construct a field with 9 elements.
3. Construct a field with 27 elements.
4. Suppose that F is a finite field of characteristic p. If c is a primitive element of F, prove that c^p is a primitive element of F.
5. Let F be a finite field of characteristic p. If $n = [F : \mathbb{Z}_p]$, prove that there exists $c \in F$ such that c is algebraic of degree n over $\mathbb{Z}_p$ and $F = \mathbb{Z}_p(c)$.
6. If F is a finite field of p^n elements, p a prime and n a positive integer, prove that $[F : \mathbb{Z}_p] = n$.
7. Describe the splitting field of $x^{3^2} - x$ over $\mathbb{Z}_3$.
8. Prove that there exists an irreducible polynomial of arbitrary degree n over $\mathbb{Z}_p$.
9. If F is a subfield of $GF(p^n)$, prove that $F \simeq GF(p^m)$, where $m|n$.
10. Show that if m and n are positive integers such that $m|n$, then $GF(p^n)$ contains a unique subfield $GF(p^m)$, $p^m - 1$ divides $p^n - 1$, whence $x^{p^m-1} - 1$ divides $x^{p^n-1} - 1$ and so $x^{p^m} - x$ divides $x^{p^n} - x$.
11. Let F be a field containing $\mathbb{Z}_p$ and $f(x)$ be a polynomial over $\mathbb{Z}_p$. If $c \in F$ is a root of $f(x)$, prove that c^p is also root of $f(x)$.
12. Let $f(x) = x^p - x - [1] \in \mathbb{Z}_p[x]$. Show that a splitting field of $f(x)$ over $\mathbb{Z}_p$ is $\mathbb{Z}_p(c)$, where c is a root of $f(x)$.
13. Let F be a field and G and H be subgroups of F^*. If G and H have order n, prove that $G = H$.
14. If F is a field such that F^* is cyclic, prove that F is finite.

References

1. Aschbacher, M. The classification of finite simple groups. *Mathematics Intelligencer*, 3(2), 59–65, 1981.
2. Barnes, W. E. *Introduction to Abstract Algebra.* Boston: D.C. Heath and Company, 1963.93.
3. Bell, E. T. *Men of Mathematics.* 2d ed. New York: Simon and Schuster, 1962.
4. Burton, D. M. *Elementary Number Theory.* Boston: Allyn & Bacon, 1980.
5. Edwards, H. M. The genesis of ideal theory. *Arch. History Exact Sci.* 23, 321–378, 1980.
6. Edwards, H. M. Dedekind's invention of ideals. In *Studies in the History of Mathematics*, E.R. Phillips, ed. The Mathematical Association of America, 1987.
7. Gillispie, C. C., ed. *Dictionary of Scientific Biography,* Vols. 1-14. New York: Charles Scribner's Sons.
8. Halmos, P. R. *Naive Set Theory.* New York: Springer Verlag, 1974.
9. Hardy, G. H., and Wright, E. M. *An Introduction to the Theory of Numbers,* 4th ed. Oxford, England. Clarendon Press, 1960.
10. Herstein, I. N. *Topics in Algebra.* 2d ed. New York: Wiley, 1975.
11. Hungerford, T. W. *Algebra.* New York: Holt, Reinhart and Winston, 1974.
12. Kleiner, I. The evolution of group theory: A brief survey. *Mathematics Magazine* 59(4), 195–215, 1986.
13. Kleiner, I. A sketch of the evolution of (noncommutative) ring theory. *L'Enseignement Mathématique*, 33, 227–267, 1987.
14. Malik, D.S., Mordeson, J.N., and Sen, M.K., Fundamentals of Abstract Algebra, McGraw Hill, 1997.
15. McCoy, N. H. *The Theory of Rings.* New York, Chelsea Publishing Company, 1973.
16. Rotman, J. J. *An Introduction to the Theory of Groups.* Iowa, Wm. C. Brown, 1988.
17. Rotman, J. J. *Galois Theory.* New York: Springer Verlag, 1990.
18. Van der Waerden, B. L. *A History of Algebra.* New York: Springer Verlag, 1985.
19. Zariski, O., and Samuel, P. *Commutative Algebra,* Vol. 1. New Jersey: D. Van Nostrand Co. Inc., 1960.

INDEX