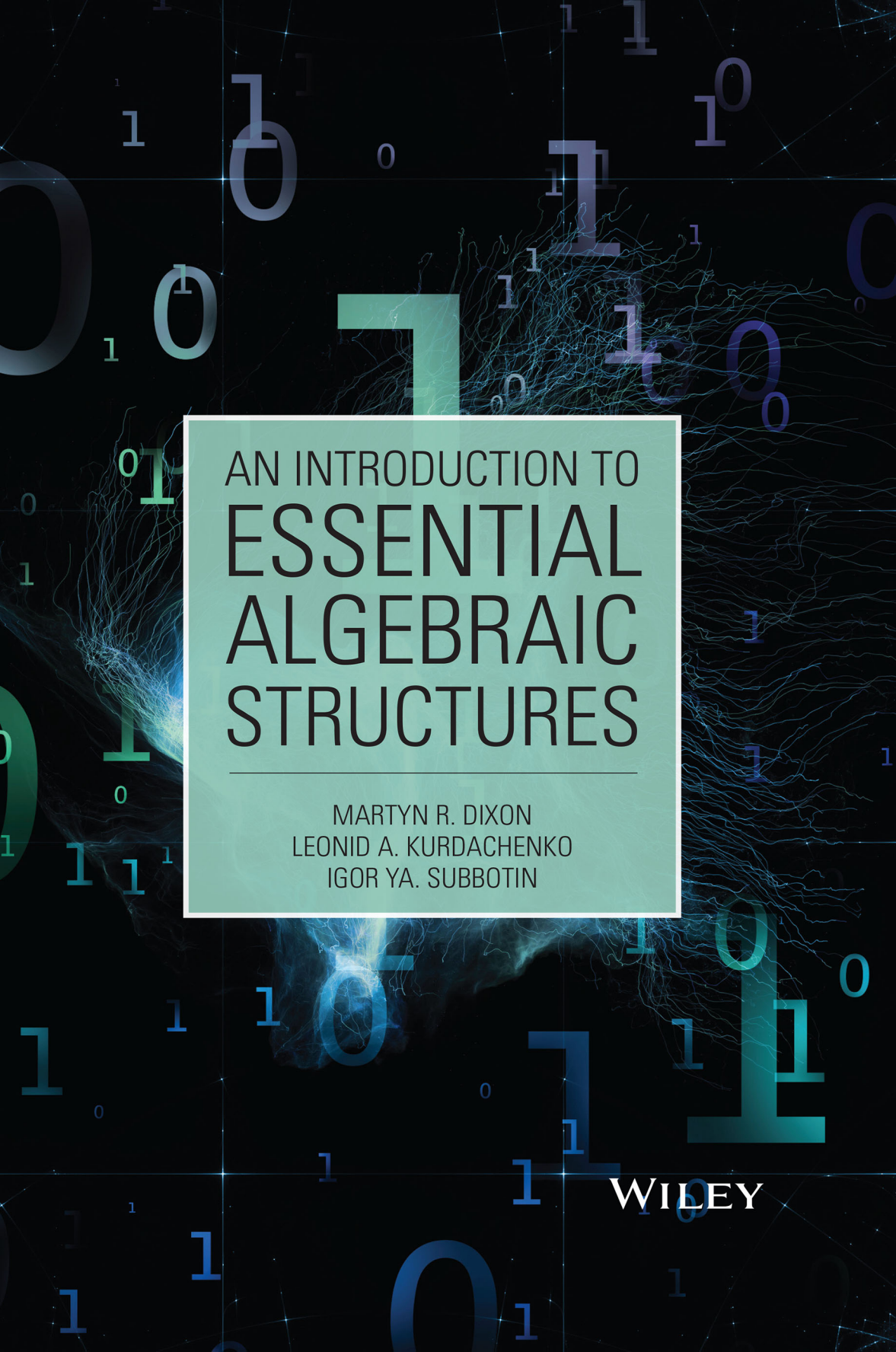

AN INTRODUCTION TO
ESSENTIAL
ALGEBRAIC
STRUCTURES
MARTYN R. DIXON
LEONID A. KURDACHENKO
IGOR YA. SUBBOTIN
WILEY

AN INTRODUCTION TO ESSENTIAL ALGEBRAIC STRUCTURES

AN INTRODUCTION TO ESSENTIAL ALGEBRAIC STRUCTURES

MARTYN R. DIXON
Department of Mathematics
The University of Alabama
Tuscaloosa, AL, USA

LEONID A. KURDACHENKO
Department of Algebra
National University of Dnepropetrovsk
Dnepropetrovsk, Ukraine

IGOR YA. SUBBOTIN
Department of Mathematics and Natural Sciences
National University
Los Angeles, CA, USA

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Dixon, Martyn R. (Martyn Russell), 1955- author.
An introduction to essential algebraic structures / Martyn R. Dixon, Department of Mathematics, The University of Alabama, Tuscaloosa, AL, Leonid A. Kurdachenko, Department of Algebra, National University of Dnepropetrovsk, Dnepropetrovsk, Ukraine, Igor Ya. Subbotin, Department of Mathematics and Natural Sciences, National University, Los Angeles, CA.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-45982-9 (cloth)
1. Ordered algebraic structures. I. Kurdachenko, Leonid A., 1949- author. II. Subbotin, Igor Ya., 1950- author. III. Title.
QA172.D59 2015
511.3′3–dc23

2014022297

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

Abstract algebra is an essential part of a mathematics program at any university. It would not be an exaggeration to say that this area is one of the most challenging and sophisticated parts of such a program. It requires beginners to establish and develop a totally different way of thinking from their previous mathematical experience. Actually, to students, this is a new language that has proved to be very effective in the investigation and description of the most important natural and mathematical laws. The transition from the well-understood ideas of Calculus, aided by its many visual examples, to the abstraction of algebra, less supported by intuition, is perhaps one of the major obstacles that students of mathematics need to overcome. Under these circumstances, it is imperative for students to have a reader-friendly introductory textbook consisting of clearly and carefully explained theoretical topics that are essential for algebra and accompanied with thoughtfully selected examples and exercises.

Abstract algebra was, until fairly recently, studied for its own sake and because it helped solve a range of mathematical questions of interest to mathematicians. It was the province of pure mathematicians. However, much of the rise of information technology and the accompanying need for computer security has its basis in abstract algebra, which in turn has ignited interest in this area. Abstract algebra is also of interest to physicists, chemists, and other scientists. There are even applications of abstract algebra to music theory. Additionally, many future high school teachers now need to have some familiarity with higher level mathematics. There is therefore a need for a growing

body of undergraduate students to have some knowledge of this beautiful subject.

We have tried to write a book that is appropriate for typical students in computer science, mathematics, mathematics education, and other disciplines. Such students should already possess a certain degree of general mathematical knowledge pertaining to typical average students at this stage. Ideally, such students should already have had a mathematics course where they have themselves written some proofs and also worked with matrices. However, the main idea of our book is that it should be as user-friendly to a beginner as possible, and for this reason, we have included material about matrices, mathematical induction, functions, and other such topics. We expect the book to be of interest not only to mathematics majors, but also to anyone who would like to learn the basic topics of modern algebra. Undergraduate students who need to take an introductory abstract algebra course will find this book very handy. We have made every effort to make the book as simple, understandable, and concise as possible, while leaving room for rigorous mathematical proofs. We illustrate the theory with a variety of examples that appeal to the previous experience of readers, which is useful in the development of an intuitive algebraic way of thinking. We cover only essential topics from the algebra curriculum typical for introductory abstract algebra courses in American universities. Through some of the numerous exercises, we introduce readers to more complex topics.

The book consists of five chapters. We start our exposition with the elements of set theory, functions, and matrix theory. In Chapter 2, we cover the main properties of the integers, viewed from an algebraic point of view. This paves the way for the final three chapters, covering Groups in Chapter 3, Rings in Chapter 4, and Fields in Chapter 5. These chapters cover the main beginning ideas of abstract algebra as well as sophisticated ideas. The book is accompanied by an Instructor's solutions manual containing solutions for all exercises in the book.

The authors would like to extend their sincere appreciation to the University of Alabama (Tuscaloosa, USA), National Dnepropetrovsk University (Dnepropetrovsk, Ukraine), and National University (Los Angeles, USA) for their great support of the authors' work. The authors also would like thank their family members for their patience, understanding, and much needed support while this work was in progress.

Martyn R. Dixon
Leonid A. Kurdachenko
Igor Ya. Subbotin

1

SETS

1.1 OPERATIONS ON SETS

The concept of a set is one of the fundamental concepts in mathematics. Set theory permeates most branches of mathematics and yet, in some way, set theory is elusive. For example, if we were to ask for the definition of a set, we may be inclined to give a response such as "it is a collection of objects" or "it is a family of things" and yet the words "collection" and "family" convey no more meaning than the word "set." The reader may be familiar with such a situation in geometry. When we talk of concepts such as points, lines, planes, and distance, we have a general idea of what we are talking about. However at some point in geometry it is necessary to have a list of axioms (the rules that we use in geometry) and definitions (of the main geometrical objects), to deduce theorems about geometry. Nevertheless, some terms must be undefined, although well-understood.

Historically, geometry was the first, best developed, theory based on a system of axioms. However, in secondary school geometry we often study geometric objects without a serious appreciation of the underlying axioms. In a similar way, set theory can also be approached somewhat informally without the kind of rigor that can be established axiomatically. In this book, this approach of using so-called "naive set theory," setting aside sophisticated

An Introduction to Essential Algebraic Structures, First Edition. Martyn R. Dixon, Leonid A. Kurdachenko and Igor Ya. Subbotin.

axiomatic constructions, is the approach we shall use. For us a set will be a collection, class, or system of well-defined and distinct objects of any nature. These objects (the *elements of the set*) are distinct, but altogether they form a new unity, a new whole—a set. We will assume that a set is defined if a rule is given or established, which allows us to determine if an object belongs to the set.

For example, we can define the set of students in the room, the set of computers connected to the Internet in the room now, the set of triangles having a right angle, the set of cars in the parking lot, and so on.

This relation of belonging is denoted by the symbol $\in$. So the fact that an element *a belongs to a set A* is denoted by $a \in A$. This is usually said "*a* is an element of *A*." If an object *b does not belong to A*, then we will write $b \notin A$. It is important to realize that for each object *a* and for each set *A we can have only one of two possible cases*, namely that $a \in A$ or $a \notin A$.

For example, if we define the set $\mathbb{N}$ to be the set of all counting (or natural) numbers, then we observe that $2 \in \mathbb{N}, 3 \in \mathbb{N}, 1034 \in \mathbb{N}$, but $-2 \notin \mathbb{N}, \frac{1}{3} \notin \mathbb{N}$, $\sqrt[3]{12} \notin \mathbb{N}$, and so on.

For a finite set *A* we can *list all its elements* (this is one way of defining a set). If the elements of *A* are denoted by $a_1, a_2, \ldots, a_n$ (here the $\ldots$ indicates that the pattern continues), then we write *A* in the following standard form,

$$A = \{a_1, a_2, \ldots, a_n\}.$$

For example, $A = \{1, 3, 5, 10\}$ means that the set *A* consists of the numbers $1, 3, 5, 10$. In such a case it is easy to see if an object belongs to this set or not. For instance, the number 1 is an element of this set, while the number 11 is not.

For **another example** let $B = \{\triangleright, \subset, \subseteq, \trianglelefteq\}$. The symbols $\triangleright, \subset, \subseteq, \trianglelefteq$ are elements of this set, but $\lhd$ is not which means $\lhd \notin B$.

We note that the element *a* and the set $\{a\}$ are different entities; here $\{a\}$ is a set, having only one element *a* (sometimes called *a singleton*). Thus the presence or absence of { and } is very important.

However, even when a set only has a finite number of elements, it is sometimes not easy to define the set by just listing its elements. The set could be very large, as is the case when we consider the set of all atoms in our pencil.

In this case, we can *assign a certain property that uniquely characterizes elements and unifies them within the given set.* This is a common way of defining a set. If $P(x)$ is some defining property that an element *x* of a set *A* either has or does not have then we use the notation

$$A = \{x \mid P(x)\}.$$

This is literally described as "the set of x such that $P(x)$." Some authors use the notation $\{x : P(x)\}$ instead.

For example, the set of all real numbers belonging to the segment $[2,5]$ is written as $\{x \mid x \in \mathbb{R} \text{ and } 2 \leq x \leq 5\}$ or as $\{x \in \mathbb{R} | 2 \leq x \leq 5\}$. Here $\mathbb{R}$ is the set of all real numbers.

It is important to note that the same set can be determined by distinct defining properties. **For example**, the set X of all solutions of the equation $x^2 - 3x + 2 = 0$, and the set Y consisting of the first two counting numbers have the same elements, namely, the numbers 1 and 2.

We use the following conventional notation for the following sets of numbers.

$\mathbb{N}$ *is the set of all natural numbers, so* $\mathbb{N} = \{1, 2, 3, \dots\}$*;*
$\mathbb{Z}$ *is the set of all integers, so* $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$*;*
$\mathbb{Q}$ *is the set of all rational numbers, so* $\mathbb{Q} = \{a/b | a, b \in \mathbb{Z}, b \neq 0\}$*;*
$\mathbb{R}$ *is the set of all real numbers.*

By common agreement, the number 0 *is not a natural number*. We write $\mathbb{N}_0$ for the set consisting of all natural numbers and the number 0 (*the set of whole numbers*).

Now we shall introduce the most important concepts related to sets.

Definition 1.1.1. *Two sets A and B are called equal if every element of A is an element of B and conversely, every element of B is an element of A. We then write* $A = B$.

A very important set is *the empty set*.

Definition 1.1.2. *A set is said to be empty if it has no elements. The empty set is denoted by* $\emptyset$.

Definition 1.1.1 shows that the empty set is unique. The empty set is always obtained if there is a contradictory property. For example, $\emptyset = \{x \mid x \in \mathbb{R} \text{ and } 2^x < 0\}$.

Definition 1.1.3. *A set A is a subset of a set B if every element of A is an element of B. This is denoted by* $A \subseteq B$.

Note that the sets *A and B are equal if and only if* $A \subseteq B$ *and* $B \subseteq A$. Indeed, in this case, every element of A is an element of B, and every element of B is an element of A. From this definition we see that the empty set is a subset of each set, and every set A is a subset of itself.

Definition 1.1.4. *A subset A of a set B is called a proper subset of B if A is a subset of B and* $A \neq B$. *This is written* $A \subset B$ *or* $A \subsetneqq B$.

In this case there exists an element $x \in A$ such that $x \notin B$. So the only subset of a nonempty set A that is not a proper subset of A is the set A itself. All other subsets of A are proper subsets of A.

Example. Let A be the set of all rectangles in the plane. Then the set B of all squares in the plane is a proper subset of A.

Again we emphasize the notation. If A is a set and a is an element of A then it is correct to write $a \in A$, but in general it will not be true that $\{a\} \in A$. However $a \in A$ if and only if $\{a\} \subseteq A$. For **example,** let A be the set of all subsets of the set $B = \{\triangleright, \subset, \subseteq\}$. Then $\{\triangleright\} \in A$, $\{\triangleright\} \subseteq B$, but of course, $\{\triangleright\} \notin B$.

Definition 1.1.5. *Let A be a set. Then the set of all subsets of A is denoted by* $\mathfrak{B}(A)$ *and is called the Boolean, or power set, of A. Thus* $\mathfrak{B}(A) = \{X \mid X \subseteq A\}$.

Example. Let $B = \{\triangleright, \subseteq, \trianglelefteq\}$. In this case

$$\mathfrak{B}(B) = \{\emptyset, \{\triangleright\}, \{\subseteq\}, \{\trianglelefteq\}, \{\triangleright, \subseteq\}, \{\triangleright, \trianglelefteq\}, \{\subseteq, \trianglelefteq\}, \{\triangleright, \subseteq, \trianglelefteq\}\}$$

is the power set of B. Notice that the set B consists of three elements, while the set $\mathfrak{B}(B)$ consists of eight elements, and that $2^3 = 8$. This is not a coincidence, but illustrates the general rule stating that if a set consists of n elements, then its power set consists of 2^n elements. This rule plays an important role in set theory and can be extended to the infinite case.

Next we introduce some operations on sets.

Definition 1.1.6. *Let A and B be sets. Then* $A \cap B$ *is the set of all elements that belong to A and to B simultaneously. This is called the intersection of A and B. Thus*

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

Example. If $A = \{1, 2, 3, 4, 5\}, B = \{3, 5, 6, 10\}$, then $A \cap B = \{3, 5\}$.

Definition 1.1.7. *Let A and B be sets. Then $A \cup B$ is the set of all elements that belong to A or to B, or both, called the union of A and B. Thus*

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

Example. If $A = \{1,2,3,4,5\}, B = \{3,5,6,10\}$, then $A \cup B = \{1,2,3,4,5,6,10\}$.

Definition 1.1.8. *Let A and B be sets. Then $A \setminus B$ is the set of all elements that belong to A but not to B, called the difference of A and B. Thus*

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

If $B \subseteq A$, then $A \setminus B$ is called the complement of B in A.

Example. Let A be the set of all right-handed people, and let B be the set of all people with brown hair.

Then:

$A \cap B$ is the set of all right-handed, brown-haired people,

$A \cup B$ is the set of all people who are right-handed or brown-haired or both,

$A \setminus B$ is the set of all people who are right-handed but not brown-haired, and

$B \setminus A$ is the set of all people who have brown hair but are not right-handed.

Example. The set of irrational numbers is the complement of the set $\mathbb{Q}$ of rational numbers in the set $\mathbb{R}$ of all real numbers.

The set $\{0\}$ is the complement of the set $\mathbb{N}$ of all natural numbers in the set $\mathbb{N}_0$ of whole numbers.

We collect together some of the standard results concerning operations on sets.

Theorem 1.1.9. *Let A, B, and C be sets.*

(i) *$A \subseteq B$ if and only if $A \cap B = A$ or $A \cup B = B$. In particular, $A \cup A = A = A \cap A$ (the idempotency of intersection and union).*

(ii) *$A \cap B = B \cap A$ and $A \cup B = B \cup A$ (the commutative property of intersection and union).*

(iii) *$A \cap (B \cap C) = (A \cap B) \cap C$ and $A \cup (B \cup C) = (A \cup B) \cup C$ (the associative property of intersection and union).*

(iv) *$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ (the distributive property).*

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

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

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

Proof. The proofs of the majority of these assertions are easy to write using the definitions. However, to indicate how the proofs may be written, we give a proof of (iv).

Let $x \in A \cap (B \cup C)$. It follows from the definition that $x \in A$ and $x \in B \cup C$. Since $x \in B \cup C$ either $x \in B$ or $x \in C$ and hence either x is an element of both sets A and B, or x is an element of both sets A and C. Thus $x \in A \cap B$ or $x \in A \cap C$, which is to say that $x \in (A \cap B) \cup (A \cap C)$. This shows that $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$.

Conversely, since $B \subseteq B \cup C$ we have $A \cap B \subseteq A \cap (B \cup C)$. Likewise $A \cap C \subseteq A \cap (B \cup C)$ and hence $(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C)$.

We can extend the notions of intersection and union to arbitrary families of sets. Let $\mathfrak{S}$ be a family of sets. Thus the elements of $\mathfrak{S}$ are also sets.

Definition 1.1.10. *The intersection of the family $\mathfrak{S}$ is the set of elements that belong to each set S from the family $\mathfrak{S}$ and is denoted by $\cap\mathfrak{S}$. Thus:*

$$\cap\mathfrak{S} = \bigcap_{S \in \mathfrak{S}} S = \{x \mid x \in S \text{ for each set } S \in \mathfrak{S}\}.$$

Definition 1.1.11. *The union of the family $\mathfrak{S}$ is the set of elements that belong to at least one set S from the family $\mathfrak{S}$ and is denoted by $\cup\mathfrak{S}$. Thus:*

$$\cup\mathfrak{S} = \bigcup_{S \in \mathfrak{S}} S = \{x \mid x \in S \text{ for some set } S \in \mathfrak{S}\}.$$

The idea of an ordered pair of real numbers is very familiar to most students of mathematics and we now extend this idea to arbitrary sets A and B. A pair of elements (a,b) where $a \in A, b \in B$, taken in the given order, is called *an ordered pair*. By definition, $(a,b) = (a_1,b_1)$ if and only if $a = a_1$ and $b = b_1$.

Definition 1.1.12. *Let A and B be sets. Then the set $A \times B$ of all ordered pairs (a,b), where $a \in A, b \in B$, is called the Cartesian product of the sets A and B. If $A = B$, then we call $A \times A$ the Cartesian square of the set A and write $A \times A$ as A^2.*

The real plane $\mathbb{R}^2$ is a natural **example** of a Cartesian product. The Cartesian product of two segments of the real number line could be interpreted geometrically as a rectangle whose sides are these segments.

Example. If $A = \{1,2,3,4\}$ and $B = \{a,b,c\}$, the Cartesian product $A \times B = \{(1,a),(1,b),(1,c),(2,a),(2,b),(2,c),(3,a),(3,b),(3,c),(4,a),(4,b),(4,c)\}$.

To make sure that none of the ordered pairs are missed, remember that if the set A consists of four elements, and the set B consists of three elements, their product must have $4 \times 3 = 12$ elements. More generally, if A has m elements and B has n elements, then $A \times B$ has mn elements.

It is easy to extend the notion of a Cartesian product of two sets to the Cartesian product of a finite family of sets.

Definition 1.1.13. *Let n be a natural number and let $A_1,\ldots,A_n$ be sets. Then the set*

$$A_1 \times \cdots \times A_n = \prod_{1 \le i \le n} A_i$$

of all ordered n-tuples $(a_1,\ldots,a_n)$ where $a_j \in A_j$, for $1 \le j \le n$, is called the Cartesian product of the sets $A_1,\ldots,A_n$.

Here $(a_1,\ldots,a_n) = (b_1,\ldots,b_n)$ if and only if $a_1 = b_1,\ldots,a_n = b_n$.

The element a_j is called *the j-th coordinate or j-th component of* $(a_1,\ldots,a_n)$.

If $A_1 = \cdots = A_n = A$ we call $\underbrace{A \times A \times \cdots \times A}_{n}$ *the n-th Cartesian power* A^n of the set A.

We shall use the convention that if A is a nonempty set then A^0 will denote a one-element set and we shall denote A^0 by $\{*\}$, where $*$ denotes the unique element of A^0. Naturally, $A^1 = A$.

Example. The most natural example of a Cartesian product of more than two sets is real three-dimensional space $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$.

We note that the commutative law is not valid in general for Cartesian products, which is to say that in general $A \times B \neq B \times A$ if $A \neq B$. The same can also be said for the associative law: It is normally the case that $A \times (B \times C)$, $(A \times B) \times C$, and $A \times B \times C$ are distinct sets.

Exercise Set 1.1

In each of the following questions explain your reasoning by giving a proof of your assertion or by using appropriate examples.

1.1.1. Which of the following assertions are valid for all sets A, B, and C?

(i) If $A \notin B$ and $B \notin C$, then $A \notin C$.

(ii) If $A \notin B$ and $B \not\subseteq C$, then $A \notin C$.

1.1.2. Which of the following assertions are valid for all sets A, B, and C?

(i) If $A \subseteq B$, $A \neq B$ and $B \subseteq C$, then $C \nsubseteq A$.

(ii) If $A \subseteq B$, $A \neq B$ and $B \in C$, then $A \notin C$.

1.1.3. Give examples of sets A, B, C, D satisfying all of the following conditions: $A \subseteq B, A \neq B, B \in C, C \in D$.

1.1.4. Give examples of sets A, B, C satisfying all the following conditions: $A \in B, B \in C$, but $A \notin C$.

1.1.5. Let

$A = \{x \in \mathbb{Z} | x = 2y \text{ for some } y \geq 0\}$;
$B = \{x \in \mathbb{Z} | x = 2y - 1 \text{ for some } y \geq 0\}$;
$C = \{x \in \mathbb{Z} |\ x < 10\}$.
Find $\mathbb{Z} \backslash A$ and $\mathbb{Z} \backslash (A \cap B)$

1.1.6. Let

$A = \{x \in \mathbb{Z} | x = 2y \text{ for some } y \geq 0\}$;
$B = \{x \in \mathbb{Z} | x = 2y - 1 \text{ for some } y \geq 0\}$;
$C = \{x \in \mathbb{Z} |\ x < 10\}$.
Find $\mathbb{Z} \backslash C$ and $C \backslash (A \cup B)$.

1.1.7. Let S be the set of all complex roots of the polynomial $f(X) \in \mathbb{R}[X]$. Suppose that $f(X) = g(X)h(X)$. Let S_1 (respectively S_2) be the set of all roots of the polynomial $g(X)$ (respectively $h(X)$). Prove that $S = S_1 \cup S_2$.

1.1.8. Let $g(X)$ and $h(X)$ be polynomials with real coefficients. Let S_1 (respectively S_2) be the set of all real roots of the polynomial $g(X)$ (respectively $h(X)$). Let S be the set of all real roots of the polynomial $f(X) = (g(X))^2 + (h(X))^2$. Prove that $S = S_1 \cap S_2$.

1.1.9. Prove that $\mathfrak{B}(A \cap B) = \mathfrak{B}(A) \cap \mathfrak{B}(B)$.

1.1.10. Prove that the equation $\mathfrak{B}(A) \cup \mathfrak{B}(B) = \mathfrak{B}(A \cup B)$ implies that either $A \subseteq B$ or $B \subseteq A$.

1.1.11. Prove that if A, B are sets then $A \setminus (A \setminus B) = A \cap B$.

1.1.12. Prove that if A, B, C are sets then $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$.

1.1.13. Let $A_n = [0, 1/n)$, for each natural number n. What is $\cap_{n \geq 1} A_n$?

1.1.14. Let $A_n = (0, 1/n]$, for each natural number n. What is $\cap_{n \geq 1} A_n$?

1.1.15. Do there exist nonempty sets A, B, C such that $A \cap B \neq \emptyset, A \cap C = \emptyset, (A \cap B) \backslash C = \emptyset$?

1.1.16. Let $A = \{1,2,3,4,5,6,7\}, B = \{2,5,7,8,9,10\}$. Find $A \cap B, A \cup B, A \setminus B, B \setminus A$, the complement of A in $\mathbb{N}$, the number of elements in $A \times B$, and the number of elements in $\mathfrak{B}(A)$.

1.1.17. Let A, B, C be sets. Prove or disprove: $(A \cap B) \times C = (A \times C) \cap (B \times C)$.

1.1.18. Let A, B, C be sets. Prove or disprove: $(A \cup B) \setminus C = (A \setminus C) \cup (B \setminus C)$.

1.1.19. The symmetric difference of two sets A, B is defined by $A \triangle B = (A \cup B) \setminus (A \cap B)$. Prove that $A \triangle B = (A \setminus B) \cup (B \setminus A)$. Also prove that $A \cap (B \triangle C) = (A \cap B) \triangle (A \cap C)$ and $A \triangle (A \triangle B) = B$.

1.1.20. Is it possible to find three sets A, B, C such that $A \cap B \neq \emptyset, A \cap C \neq \emptyset, B \cap C \neq \emptyset$, but $A \cap B \cap C = \emptyset$.

1.2 SET MAPPINGS

The notion of a mapping (or function) plays a key role in mathematics.

A *mapping (or a function)* f from a set A to a set B is defined if for each element of A there is a rule that associates a uniquely determined element of B. This is usually written $f : A \longrightarrow B$. If $a \in A$, then the unique element $b \in B$, which corresponds to a, is denoted by $f(a)$ and we sometimes write $a \longmapsto b$. We say that $b = f(a)$ is an *image* of a, and a is a *preimage or inverse image* of b. Each element of A has one and only one image. However, an element $b \in B$ can have several preimages or no preimage at all. If $b \in B$, then we denote the set of preimages of b by $f^{-1}(b) = \{a \in A | f(a) = b\}$. Of course $f^{-1}(b) = \emptyset$ if there are no preimages of b.

The set A is called the *domain* of the mapping f, while the set of all images of all elements of A is a subset of B called the *range* of f which we denote by $\mathbf{Im}(f)$. The set B is usually called the *codomain* of f.

Example. Functions should look familiar to you since you already worked with them in Calculus courses. You can look at the function $y = x^2$ defined on the set $\mathbb{R}$ of real numbers as a mapping of the set $\mathbb{R}$ of real numbers to itself. Here the law of association is the unique number $y = x^2$, corresponding to each $x \in \mathbb{R}$. In this case the domain is $A = \mathbb{R}$, and the range is $B = \mathbf{Im}(f) = \mathbb{R}^+$ the set of all nonnegative real numbers.

We consider a further example having no relation to Calculus.

Example. Let A be the set of all people, B (respectively C) be the set of all males (respectively all females). Define the function $m : A \longrightarrow B$ (respectively $w : A \longrightarrow C$) by the rule that, for each person a, the image $m(a)$ is his/her father (respectively $w(a)$ is his/her mother).

A function can be thought of as a *correspondence* between sets A and B and in particular as a set of ordered pairs (a,b) where the first element a of the pair belongs to the first set A, the domain, and the second element b of the pair belongs to the second set B, the corresponding codomain. Note, that in terms of the Cartesian product, a correspondence between A and B is a subset of $A \times B$.

If $A = B$, then we will say there is a correspondence or a *relation* (or more precisely a *binary relation*) between the elements of A.

Example. Define the mapping $f : A \longrightarrow B$, where $A = \{-1,2,3,5\}$, $B = \{0,2,8,9,11\}$ by the rule: $-1 \longmapsto 2$, $2 \longmapsto 0$, $3 \longmapsto 9$, $5 \longmapsto 8$. The mapping f is defined as the set $\{(-1,2),(2,0),(3,9),(5,8)\}$ and we write $f(-1) = 2, f(2) = 0$ and so on.

As you can see, not all elements from B are involved: $11 \in B$ does not have a match in A. Thus $f^{-1}(11) = \emptyset$.

Thus, with each function $f : A \longrightarrow B$ we can form the set $\{(x,f(x)) \mid x \in A\} \subseteq A \times B$. This subset is called *the graph* of the function f.

Note that not every correspondence can serve as the graph of a function. Only a set of ordered pairs in which each element of the domain has only *one* element associated with it in the range is the graph of a function.

For example, the set of ordered pairs $\{(-1,2),(2,0),(3,9),(5,8),(-1,11)\}$ defines a correspondence between the sets $A = \{-1,2,3,5\}$ and $B = \{0,2,8,9,11\}$ but this does not correspond to a function since the element $-1 \in A$ is connected with two elements, 11 and 2, of B.

For further **examples**, let Φ denote the relation on the set of all people where $(a,b) \in \Phi$ means that a and b are people who have cars of the same brand (let's say Mercedes). This relation will not be a function.

Next, let A be the set of all points in a plane, and let B be the set of all lines in this plane. Let Γ be the correspondence defined by $(a,b) \in \Gamma$, if the point a belongs to the line b. Again this is not a function, since a given point will lie on many lines.

Definition 1.2.1. *The functions $f : A \longrightarrow B$ and $g : C \longrightarrow D$ are said to be equal if $A = C, B = D$ and $f(a) = g(a)$ for each element $a \in A$.*

Some useful and common terminology can be found in the following definition.

Definition 1.2.2. *Let $f : A \longrightarrow B$ be a function.*

(i) *The function f is said to be injective (or one-to-one) if every pair of distinct elements of A have distinct images.*

(ii) *The function f is said to be surjective (or onto) if* $\mathbf{Im} f = B$.

(iii) *The function f is said to be bijective if it is injective and surjective. In this case f is a one-to-one, onto correspondence.*

Examples. First, observe that the function $f : \mathbb{R} \longrightarrow \mathbb{R}$, satisfying $f(x) = x^2$ for all $x \in \mathbb{R}$, is neither injective nor surjective, since, for example, $f(-2) = f(2) = 4$ and $-1 \neq x^2$, for all $x \in \mathbb{R}$. However, the function $f_1 : \mathbb{R}^+ \longrightarrow \mathbb{R}$, satisfying $f_1(x) = x^2$ for all $x \in \mathbb{R}^+$, is injective and the function $f_2 : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$, satisfying $f_2(x) = x^2$ for all $x \in \mathbb{R}^+$, is injective and surjective (so is bijective).

Another familiar example is the function $y = \ln x$. Here we have a bijective mapping from the set $\mathbb{R}_+ \backslash \{0\}$ of all positive real numbers to the set $\mathbb{R}$ of all real numbers.

The correspondence $\{(-1,2),(2,0),(3,9),(5,8)\}$ considered above, from the set $A = \{-1,2,3,5\}$ into the set $B = \{0,2,8,9,11\}$ is injective, but will only be bijective if we delete 11 from the set B.

The function $m : A \longrightarrow B$ from the set A of all people to the set B of all males, where $m(a)$ is the father of person a, is not injective and not surjective.

The following statement is immediate from the definitions.

Proposition 1.2.3. *Let $f : A \longrightarrow B$ be a function. Then*

(i) *f is injective if and only if every element of B has at most one preimage;*

(ii) *f is surjective if and only if every element of B has at least one preimage;*

(iii) *f is bijective if and only if every element of B has exactly one preimage.*

We say that a set A *is finite* if there is a positive integer n, for which there exists a bijective mapping $A \longrightarrow \{1,2,\ldots,n\}$. Thus we can count the elements of A and the positive integer n is called the order of the set A; we will write this as $|A| = n$ or **Card** $A = n$. The empty set is finite and its order is 0. A set that is not finite is called *infinite*. The following assertions are also easy to see.

Corollary 1.2.4. *Let A and B be finite sets and let $f : A \longrightarrow B$ be a mapping.*

(i) *If f is injective, then $|A| \leq |B|$;*

(ii) *If f is surjective, then $|A| \geq |B|$;*

(iii) *If f is bijective, then $|A| = |B|$.*

The next result is more interesting.

Corollary 1.2.5. *Let A be a finite set and let $f : A \longrightarrow A$ be a mapping.*

(i) *If f is injective, then f is bijective;*
(ii) *If f is surjective, then f is bijective.*

Proof.

(i) If f is injective it is evident from Corollary 1.2.4 that $|A| \leq |\mathbf{Im}(f)|$. Since also $|\mathbf{Im}(f)| \leq |A|$, it follows that $|A| = |\mathbf{Im}(f)|$. Since $\mathbf{Im}(f)$ is a subset of A, this equation implies $A = \mathbf{Im}(f)$, so f is surjective and hence bijective.
(ii) Now suppose that f is surjective. If $a, b \in A$, $a \neq b$ and $f(a) = f(b)$ then the map $g : A \setminus \{a\} \longrightarrow A$ is also surjective and by Corollary 1.2.4 it follows that $|A \setminus \{a\}| \geq |A|$, which is a contradiction. Hence $a = b$ and f is injective.

Definition 1.2.6. *Let A be a set. The mapping $\varepsilon_A : A \longrightarrow A$, defined by $\varepsilon_A(a) = a$, for each $a \in A$, is called the identity mapping of A.*

If C is a subset of A, then the mapping $j_C : C \longrightarrow A$, defined by $j_C(c) = c$ for each element $c \in C$, is called a canonical injection or an identical embedding.

Definition 1.2.7. *Let $f : A \longrightarrow B$ and $g : C \longrightarrow D$ be mappings. Then we say that f is the restriction of g, or g is an extension of f, if $A \subseteq C, B \subseteq D$ and $f(a) = g(a)$ for each element $a \in A$.*

For example, a canonical injection is the restriction of the corresponding identity mapping. Note that a restriction of g is uniquely defined once the subsets A and B have been specified; however there are many different extensions of a mapping. We introduce our next topic rather informally.

Definition 1.2.8. *A set A is called countable if there exists a bijective mapping $f : \mathbb{N} \longrightarrow A$.*

In the case when A is countable, we often write $a_n = f(n)$ for each $n \in \mathbb{N}$. Then

$$A = \{a_1, a_2, \ldots, a_n, \ldots\} = \{a_n \mid n \in \mathbb{N}\}.$$

Thus the elements of a countable set can be indexed (or numbered) by the set of all positive integers. Conversely, if all elements of an infinite set A can be indexed using natural numbers, then A is countable. The bijection from $\mathbb{N}$ to A here is natural: every natural number n corresponds to the element a_n of A with index n.

Proposition 1.2.9. *Let A and B be sets and suppose that A is countable. If there exists a bijective mapping $f : B \longrightarrow A$ (respectively $g : A \longrightarrow B$), then B is countable.*

Proof. Since A is countable, we can write $A = \{a_n | n \in \mathbb{N}\}$.

First suppose that there is a bijection $g : A \longrightarrow B$. Consider the mapping $g_1 : \mathbb{N} \longrightarrow B$, defined by $g_1(n) = g(a_n), n \in \mathbb{N}$. It is easy to see that g_1 is bijective.

Suppose now that there exists a bijection $f : B \longrightarrow A$. Then each element a of A has exactly one preimage $f^{-1}(a)$. Consider the mapping $f_1 : \mathbb{N} \longrightarrow B$ defined by $f_1(n) = f^{-1}(a_n), n \in \mathbb{N}$. It is easy to see that f_1 is bijective.

Theorem 1.2.10.

(i) *Let A be an infinite set. Then A contains a countable subset;*
(ii) *Let A be a countable set and let B be a subset of A. If B is infinite, then B is countable;*

Proof.

(i) Since A is infinite it is not empty so choose $a_1 \in A$. The subset $A \setminus \{a_1\}$ is also not empty, therefore we can choose an element a_2 in this subset. Since A is infinite, $A \setminus \{a_1, a_2\} \neq \emptyset$, so that we can choose an element a_3 in this subset and so on. This process cannot terminate after finitely many steps because A is infinite. Hence A contains the infinite subset $\{a_n \mid n \in \mathbb{N}\}$, which is countable.
(ii) Let $A = \{a_n \mid n \in \mathbb{N}\}$. Then there is a least positive integer $k(1)$ such that $a_{k(1)} \in B$ and we put $b_1 = a_{k(1)}$. There is a least positive integer $k(2)$ such that $a_{k(2)} \in B \setminus \{b_1\}$. Put $b_2 = a_{k(2)}$, and so on. This process cannot terminate since B is infinite. Then all the elements of B will be indexed by positive integers.

Notice that Theorem 1.2.10 implies that a subset of a countable set is either countable or finite.

Corollary 1.2.11. *Every infinite subset of $\mathbb{N}$ is countable.*

Corollary 1.2.12. *Let A and B be sets. If A is countable and there is an injective mapping $f : B \longrightarrow A$, then B is finite or countable.*

Proof. We consider the mapping $f_1 : B \longrightarrow \mathbf{Im}f$ defined by $f_1(b) = f(b)$, for each element $b \in B$. By this choice, f_1 is surjective. Since f is injective, f_1 is

also injective and hence f_1 is bijective. Finally, Theorem 1.2.10 implies that $\mathbf{Im}f$ is finite or countable.

Example. The set of integers is countable. We can construct a bijective mapping $f : \mathbb{N} \longrightarrow \mathbb{Z}$, informally, in the following way

$$\begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & \dots \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \dots . \\ 0 & 1 & -1 & 2 & -2 & 3 & -3 & \dots \end{array}$$

It is a little bit trickier to find a bijective mapping between $\mathbb{N}$ and the set of all rational numbers $\mathbb{Q}$ and we here indicate informally how to do this. Put $\mathbb{Q}_n = \{\frac{m}{k} \mid |m| + |k| = n\}$, for each $n \in \mathbb{N}$. Then each subset $\mathbb{Q}_n$ is finite and the elements of $\mathbb{Q}_n$ can be ordered in their natural order. Now we construct a bijective mapping $r : \mathbb{N} \longrightarrow \mathbb{Q}$. We have $\mathbb{Q}_1 = \{\frac{0}{\pm 1} = 0\}$, so put $r(1) = 0$. Further, $\mathbb{Q}_2 = \{-1 = \frac{-1}{1} = \frac{1}{-1}, \frac{0}{\pm 2} = 0, 1 = \frac{1}{1} = \frac{-1}{-1}\}$. Since 0 already has a preimage, put $r(2) = -1$, $r(3) = 1$. For the next step we consider $\mathbb{Q}_3 = \{-2 = \frac{-2}{1} = \frac{2}{-1}, \frac{-1}{2} = \frac{-1}{2} = \frac{1}{-2}, \frac{0}{\pm 3} = 0, \frac{1}{2} = \frac{1}{2} = \frac{-1}{-2}, 2 = \frac{2}{1} = \frac{-2}{-1}\}$.

Again 0 already has a preimage, so put $r(4) = -2, r(5) = \frac{-1}{2}, r(6) = \frac{1}{2}, r(7) = 2$. Consider next $\mathbb{Q}_4 = \{-3 = \frac{-3}{1} = \frac{3}{-1}, \frac{-1}{3} = \frac{-1}{3} = \frac{1}{-3}, -1 = \frac{-2}{2} = \frac{2}{-2}, 0 = \frac{0}{\pm 4}, \frac{1}{3} = \frac{-1}{-3}, 1 = \frac{2}{2} = \frac{-2}{-2}, 3 = \frac{3}{1} = \frac{-3}{-1}\}$. The numbers $0, -1, 1$ have preimages, thus we need to index the numbers $-3, -\frac{1}{3}, \frac{1}{3}, 3$, so put $r(8) = -3, r(9) = \frac{-1}{3}, r(10) = \frac{1}{3}, r(11) = 3$. If we continue this process we will index all rational numbers using natural numbers.

An important natural question arises: Is there an infinite set that is not countable? The answer to this question is yes and was obtained by Georg Cantor who proved that *the set* $[0, 1]$*, and therefore the set of all real numbers, is not countable.* We shall not pursue this topic further here.

As we saw here, to establish that two sets have the same number of elements there is no need to count these elements. It is sufficient to establish the existence of a bijective mapping between these sets. This idea is really at the heart of the abstract notion of a number. By extending this to arbitrary sets we arrive at the concept of the *cardinality* of a set.

Definition 1.2.13. *Two sets A and B are called equipollent, if there exists a bijective mapping $f : A \longrightarrow B$.*

We will denote this by $|A| = |B|$.

If A and B are finite sets, then A and B are equipollent precisely when these sets have the same number of elements. More generally when two sets are equipollent we say that they have the same cardinal number. This allows us to establish an ordering of the set of cardinal numbers, but we refrain from pursuing this topic.

Exercise Set 1.2

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

1.2.1. Let $\Phi = \{(x,y) \in \mathbb{N} \times \mathbb{N} \mid 3x = y\}$. Is Φ a function?

1.2.2. Let $\Phi = \{(x,y) \in \mathbb{N} \times \mathbb{N} \mid 3x = 5y\}$. Is Φ a function?

1.2.3. Let $\Phi = \{(x,y) \in \mathbb{N} \times \mathbb{N} \mid x^2 = y^2\}$. Is Φ a function?

1.2.4. Let $\Phi = \{(x,y) \in \mathbb{N} \times \mathbb{N} \mid x = y^4\}$. Is Φ a function?

1.2.5. Let $f : \mathbb{Z} \longrightarrow \mathbb{N}_0$ be the mapping defined by $f(n) = |n|$, where $n \in \mathbb{Z}$. Is f injective? Is f surjective?

1.2.6. Let $f : \mathbb{N} \longrightarrow \{x \in \mathbb{Q} \mid x > 0\}$ be the mapping defined by $f(n) = \frac{n}{n+1}$, where $n \in \mathbb{N}$. Is f injective? Is f surjective?

1.2.7. Let $f : \mathbb{N} \longrightarrow \mathbb{N}$ be the mapping, defined by $f(n) = (n+1)^2$, where $n \in \mathbb{N}$. Is f injective? Is f surjective?

1.2.8. Let $f : \mathbb{N} \longrightarrow \mathbb{N}$ be the mapping, defined by $f(n) = \frac{n^2+n}{2}$, where $n \in \mathbb{N}$. Is f injective? Is f surjective?

1.2.9. Let $f : \mathbb{Z} \longrightarrow \mathbb{Z} \times \mathbb{Z}$ be the mapping, defined by the rule $f(n) = (n+1, n)$, where $n \in \mathbb{Z}$. Is f injective? Is f surjective?

1.2.10. Let $f : \mathbb{Z} \longrightarrow \mathbb{Z} \times \mathbb{Z}$ be the mapping, defined by $f(n) = (n, n^4)$, where $n \in \mathbb{Z}$. Is f injective? Is f surjective?

1.2.11. Let $f : \mathbb{Q} \longrightarrow \mathbb{R}$ be the map defined by $f(a) = a$, for all $a \in \mathbb{Q}$. Define $g_1 : \mathbb{R} \longrightarrow \mathbb{R}$ by $g_1(a) = a$, for all $a \in \mathbb{R}$ and define $g_2 : \mathbb{R} \longrightarrow \mathbb{R}$ by

$$g_2(a) = \begin{cases} a, \text{ if } a \in \mathbb{Q} \\ 1, \text{ if } a \notin \mathbb{Q} \end{cases}$$

Show that g_1, g_2 are both extensions of f.

1.2.12. Let A and B be finite sets, with $|A| = a, |B| = b$. Find the number of injective mappings from A to B.

1.2.13. Let $f : A \longrightarrow B$ be a function from the set A to the set B and let U, V be subsets of A. Give a proof or counterexample to the statement: $f(U \cap V) = f(U) \cap f(V)$.

1.2.14. Let A be a finite set. Prove that if $f : A \longrightarrow A$ is an injective function then f is also surjective.

1.2.15. Let $f : A \longrightarrow B$ be a function from the set A to the set B and let U, V be subsets of B. Give a proof or counterexample to the statement: $f^{-1}(U) \cap f^{-1}(V) = f^{-1}(U \cap V)$.

1.2.16. Let $f : \mathbb{N}_0 \longrightarrow \mathbb{N}_0$ be the mapping defined by $f(n) = n^2 + 3n$, where $n \in \mathbb{N}_0$. Is f injective? Is f surjective?

1.2.17. Let $f : A \longrightarrow B$ be a function from the set A to the set B and let U, V be subsets of B. Give a proof or counterexample to the statement: $f^{-1}(U) \cup f^{-1}(V) = f^{-1}(U \cup V)$.

1.2.18. Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a bijection. Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be the map defined by $g(x) = f(5x+3)$. Is g injective? Is g surjective?

1.2.19. Prove that a countable union of countable sets is again countable.

1.2.20. Prove that if A and B are countable sets then $A \times B$ is also countable.

1.3 PRODUCTS OF MAPPINGS AND PERMUTATIONS

We next consider the product of two mappings. This product will allow us to construct new mappings based on given ones, but it is not defined in all cases. If $f : A \longrightarrow B$ and $g : C \longrightarrow D$ are mappings, then the product of g and f is defined only when $B = C$.

Definition 1.3.1. *Let $f : A \longrightarrow B$ and $g : B \longrightarrow C$ be mappings. The mapping $g \circ f : A \longrightarrow C$, defined by the rule*

$$g \circ f(a) = g((f(a)) \text{ for each } a \in A$$

is called the product or the composite of g and f.

We think of this as follows. First the mapping f acts on the element $a \in A$, and then the mapping g acts on the image $f(a)$ (the result of the first mapping f applied to a). Thus, when we write $g \circ f$, the mapping f is done first, opposite to the usual rules for reading. There will be one important exception to this general rule, which we will discuss later.

Example. We are familiar with composition of real functions, so there are many standard examples that can be used to illustrate the product of functions. For example, let $f : \mathbb{R} \longrightarrow \mathbb{R}$ and $g : \mathbb{R} \longrightarrow \mathbb{R}$ be defined by $f(x) = 4x - 1$, $g(x) = x^2 + 1$, for all $x \in \mathbb{R}$. In this case, both products $g \circ f$ and $f \circ g$ are defined and we now evaluate these compositions, using two slightly different methods. For every element $x \in \mathbb{R}$ we have

$$(g \circ f)(x) = g(f(x)) = f(x)^2 + 1 = (4x-1)^2 + 1 = 16x^2 - 8x + 2, \text{ and}$$
$$(f \circ g)(x) = f(g(x)) = f(x^2+1) = 4(x^2+1) - 1 = 4x^2 + 3$$

We next consider an example of the product of two nonnumeric functions. Let A be the set of all people and consider the functions $m : A \longrightarrow A$ and $w : A \longrightarrow A$, where $m(a)$ is the father of person a, and $w(a)$ is the mother of person a. In this case, both products $m \circ w$ and $w \circ m$ are defined. Then $(m \circ w)(a) = m(w(a))$ is the father of the mother of the person a, which is the grandfather on the mother's side, while $(w \circ m)(a) = w(m(a))$ is the mother of the father of the person a, which is the grandmother on the father's side.

These examples show that the product of mappings is not a commutative operation. In general, the situation when $g \circ f = f \circ g$ is fairly rare. We say that the mappings $f : A \longrightarrow B$ and $g : C \longrightarrow D$ *permute or commute* if both products $g \circ f$ and $f \circ g$ exist (i.e., $C = B$ and $D = A$) and $g \circ f = f \circ g$, in which case $A = B = C = D$.

However, function composition always satisfies the associative property, at least when the products are defined.

Theorem 1.3.2. *Let $f : A \longrightarrow B, g : B \longrightarrow C$ and $h : C \longrightarrow D$ be functions. Then $h \circ (g \circ f) = (h \circ g) \circ f$.*

Proof. We have $g \circ f : A \longrightarrow C, h \circ g : B \longrightarrow D$ and so

$$h \circ (g \circ f) : A \longrightarrow D, (h \circ g) \circ f : A \longrightarrow D.$$

If a is an arbitrary element of A, then

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

whereas

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

Hence $\big(h \circ (g \circ f)\big)(a) = \big((h \circ g) \circ f\big)(a)$ for all $a \in A$ which proves that $(h \circ g) \circ f = h \circ (g \circ f)$.

Let $f : A \longrightarrow B$ be a mapping. It is not hard to see that

$$\varepsilon_B \circ f = f \circ \varepsilon_A = f,$$

so the mappings ε_B and ε_A play the role of "left identity" and "right identity" elements, respectively, for the operation of multiplication of mappings. Also it should be noted that there is no "universal" identity element for all mappings.

Definition 1.3.3. *Let $f : A \longrightarrow B$ be a mapping. Then the mapping $g : B \longrightarrow A$ is called an inverse of f if $g \circ f = \varepsilon_A$ and $f \circ g = \varepsilon_B$.*

We remark first that if f has an inverse then it is unique. To show this let $g : B \longrightarrow A$ and $h : B \longrightarrow A$ be mappings satisfying

$$g \circ f = \varepsilon_A, f \circ g = \varepsilon_B \text{ and } h \circ f = \varepsilon_A, f \circ h = \varepsilon_B.$$

Now consider the product $g \circ f \circ h$. We have

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

Theorem 1.3.4. *Let $f : A \longrightarrow B, g : B \longrightarrow A$ be mappings. If $g \circ f = \varepsilon_A$, then f is an injective mapping and g is a surjective mapping.*

Proof. Suppose that A has elements a and c such that $f(a) = f(c)$. Then

$$a = \varepsilon_A(a) = (g \circ f)(a) = g(f(a)) = g(f(c)) = g \circ f(c) = \varepsilon_A(c) = c,$$

which shows that f is injective.

Next, let u be an arbitrary element of A. Then

$$u = \varepsilon_A(u) = g \circ f(u) = g(f(u)),$$

and, in particular, $f(u)$ is a preimage of the element u relative to g. It follows that $\mathbf{Im}\ g = A$, so g is surjective.

Corollary 1.3.5. *Let $f : A \longrightarrow B$ be a mapping. Then f has an inverse mapping if and only if f is bijective. In this case, the inverse mapping is also bijective.*

Proof. Suppose that f has inverse mapping $g : B \longrightarrow A$. Then $g \circ f = \varepsilon_A$ and $f \circ g = \varepsilon_B$. From the first equation and Theorem 1.3.4 it follows that f is injective and g is surjective. Applying Theorem 1.3.4 to the second equation, we deduce that g is injective and f is surjective. It follows that f and g are both bijective.

Conversely, let f be bijective. By Proposition 1.2.3, every element $b \in B$ has exactly one preimage a_b. Thus $f(a_b) = b$ and we may define the mapping $g : B \longrightarrow A$ by $g(b) = a_b$. We show that g is the desired inverse to f. Indeed, if

$b \in B$ then $f(g(b)) = f(a_b) = b$ so $f \circ g = \varepsilon_B$. On the other hand, if $a \in A$ then $f(a)$ has the unique preimage a and so $g(f(a)) = a$ by definition of g. Thus $g \circ f = \varepsilon_A$ and the proof is complete.

Since a bijective mapping f has only one inverse, we use the notation f^{-1} for it. The reader is cautioned that the notation does not mean $1/f(x)$. We note also that by Corollary 1.3.5 the mapping f^{-1} is also bijective. We observe that Corollary 1.3.5 not only proves the existence of the inverse mapping, but also shows how to find it.

For **example**, consider the real functions f and g defined as follows:

$$f(x) = 4x - 1, g(x) = 5x^3 + 1 \text{ for each } x \in \mathbb{R}.$$

Once we know that f, g are bijections we can find their inverses, as usual, by "solving for x in terms of y" and often to see that a function is surjective amounts to doing just that. For example, let $b \in \mathbb{R}$. We find its unique preimage, c, relative to f by solving the equation $b = 4c - 1$. Clearly $c = \frac{b+1}{4}$ and hence $f^{-1} : \mathbb{R} \longrightarrow \mathbb{R}$ is defined as follows:

$$f^{-1}(x) = \frac{x+1}{4}.$$

Of course, we have not shown that f is injective here.

To show that g is injective we have to show that if $5x_1^3 + 1 = 5x_2^3 + 1$ then $x_1 = x_2$; however, this follows since the first equation implies that $x_1^3 = x_2^3$ so $x_1 = x_2$. Furthermore, if $b = 5a^3 + 1$ then we can solve uniquely for a to obtain $a = \sqrt[3]{\frac{b-1}{5}}$ so that the inverse of g is

$$g^{-1}(x) = \sqrt[3]{\frac{x-1}{5}}.$$

We note one further important property of the product of functions that we have already used.

Proposition 1.3.6. *Let $f : A \longrightarrow B, g : B \longrightarrow C$ be mappings.*

(i) *If f and g are injective, then $g \circ f$ is injective;*
(ii) *If f and g are surjective, then $g \circ f$ is surjective;*
(iii) *If f and g are bijective, then $g \circ f$ is bijective.*

This statement can be proved directly using the definitions.

Definition 1.3.7. *Let A be a set. A mapping from A to A is called a transformation of the set A. The set of all transformations of A is denoted by* $\mathbf{P}(A)$ *or* A^A.

We note that a product of two transformation of A is always defined and is again a transformation. Clearly, multiplication of transformations is associative and in this case there exists an identity element, namely the identity transformation ε_A.

Examples. The following transformations play a significant role in geometry. Let a be a line in space and for each point $P \in \mathbb{R}^3$ let Q be the point obtained by rotating P about the line a through an angle α. Thus a acts as the axis of rotation and this defines a transformation of $\mathbb{R}^3$ called the rotation of $\mathbb{R}^3$ about the axis a through angle α.

Another important transformation of the space $\mathbb{R}^3$ is a translation by a given vector. Of course we can consider similar transformations of the plane $\mathbb{R}^2$.

The mapping $f : \mathbb{Z} \longrightarrow \mathbb{Z}$ defined by $f(k) = k^2 + 1$, where $k \in \mathbb{Z}$, is a transformation of the set of integers $\mathbb{Z}$. It is not bijective, therefore it has no inverse. The mapping $g : \mathbb{Z} \longrightarrow \mathbb{Z}$, defined by $g(k) = -k$, where $k \in \mathbb{Z}$ is a bijective transformation of $\mathbb{Z}$, and this transformation is clearly its own inverse.

There are other important transformations. For example, it is well-known that a projection of space onto a plane is an important transformation. This transformation is not bijective since it is not one-to-one.

Bijective transformations play a particularly important role.

Definition 1.3.8. *Let A be a set. A bijective transformation of A is called a permutation of A. The set of all permutations of A is denoted by* $\mathbf{S}(A)$. *Thus* $\phi \in \mathbf{S}(A)$ *if and only if* $\phi : A \longrightarrow A$ *is a bijective mapping.*

The word "permutation" has an alternative meaning since it is also widely used in combinatorics giving us a situation when the same word represents two different things, possibly leading to ambiguity. However these two ideas are closely connected and usually it is clear from the context which meaning of the term permutation is being used.

Let A be a finite set, say $A = \{a_1, a_2, \ldots, a_n\}$. Here, the order of the elements is not important. If π is a permutation of the set A, it can be represented in the following way

$$\begin{array}{cccccc} a_1 & a_2 & a_3 & \ldots & a_{n-1} & a_n \\ \downarrow & \downarrow & \downarrow & \ldots & \downarrow & \downarrow \\ \pi(a_1) & \pi(a_2) & \pi(a_3) & \ldots & \pi(a_{n-1}) & \pi(a_n) \end{array}$$

This can be considered as a renumeration of the elements of A.

Example. Let $A = \{1,2,3,4\}$ and consider the permutation π given by the chart below

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ 4 & 3 & 2 & 1 \end{array}.$$

Now consider the set $\bar{A} = \{a_1, a_2, a_3, a_4\}$ and define the permutation $\bar{\pi}$ on $\bar{A}$ by the chart

$$\begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ a_4 & a_3 & a_2 & a_1 \end{array}.$$

Evidently, the first chart π could be used to represent the transformation $\bar{\pi}$ of the set $\bar{A}$ so we can represent a transformation of a set by indexing its elements and then tracking the changes in this indexing generated by the transformation.

In the same way, a permutation of any indexed set can be represented by the corresponding change in the indices. Since any finite set can be indexed, this gives us an easy way to represent such transformations. In this case the order of the elements in the set is important and is defined by the indexing. We shall denote a permutation of the finite set $\bar{A} = \{a_1, a_2, \ldots, a_n\}$ by an ordered tuple consisting of all the elements of $\bar{A}$ once and only once. We also say that this is a permutation of the elements $a_1, a_2, \ldots, a_n$. The elements in a tuple appear in some order: the tuple has a first element (unless it is empty), a second element (unless its length is less than 2), and so on. For example, if $\bar{A} = \{1,2,3\}$, then $(1,2,3)$ and $(3,2,1)$ are two different ways to list the elements of A in some order. These give two permutations of the numbers $1,2,3$.

Let $A = \{a_1, a_2, \ldots, a_n\}$ be a finite set with n elements, and let π be a permutation of the set A. Based on the considerations above, permutations of the set $\bar{A} = \{1,2,\ldots,n\}$ can be considered instead of permutations of the abstract set $A = \{a_1, a_2, \ldots, a_n\}$. Earlier we used the notation $\mathbf{S}(A)$ for the set of permutations of A. However, we will use the notation $\mathbf{S}_n$, or $\mathbf{Sym}(n)$, for the set of all permutations of the set $\{1,2,\ldots,n\}$. If $\pi \in \mathbf{S}_n$, then we will say that π is *a permutation of degree n*. The number of different permutations of the elements of the set A consisting of n elements is easily seen to be equal to $n! = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$. Hence $|\mathbf{S}_n| = n!$

The permutation $\pi : \{1,2,\ldots,n\} \longrightarrow \{1,2,\ldots,n\}$ can be written as

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ \pi(1) & \pi(2) & \ldots & \pi(n) \end{pmatrix}$$

which we will call the *tabular form of the permutation*. Since π is a permutation of the set $\{1,2,\ldots,n\}$ we see that

$$\{1,2,\ldots,n\} = \{\pi(1),\pi(2),\ldots,\pi(n)\}.$$

Thus the second row of a tabular form is a permutation of the numbers $1,2,\ldots,n$. It is not necessary to write all elements of the first row in the natural order from 1 to n, although this is usually the way such permutations are written. Sometimes it is convenient to write the first row in a different order. What is most important is that every element of the second row is the image of the corresponding element of the first row situated just above.

For example,

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 4 & 9 & 1 & 7 & 8 & 3 & 5 & 2 & 6 \end{pmatrix} \text{ and } \begin{pmatrix} 2 & 5 & 7 & 1 & 9 & 3 & 6 & 4 & 8 \\ 9 & 8 & 5 & 4 & 6 & 1 & 3 & 7 & 2 \end{pmatrix}$$

are the same permutation. Perhaps, for beginners, in order to better understand permutations, it may be worthwhile to write the permuation with arrows connecting the element of the first row with its image in the second row as in $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 4 & 9 & 1 & 7 & 8 & 3 & 5 & 2 & 6 \end{pmatrix}$. This method of writing a permutation should be quickly learned and then the student should revert to the shorthand notation.

We will multiply permutations π and σ by using the general rule of multiplication of mappings, namely composition of functions, introduced earlier in this section, with one important modification, suggested earlier. We note that normally we write and read from left to right. Thus, in writing the product $\pi \circ \sigma$ we first write π and then σ. Thus it is entirely natural that the first permutation to act should be the one that is written first, and after that the permutation that acts second is written and so on. Thus **for permutations only** when we write $\pi \circ \sigma$ we will mean that first the permutation π is performed and then the permutation σ. We remark that this is a personal preference and that in some books $\pi \circ \sigma$ means that first σ is performed and then π. This slight inconsistency is the result of writing mappings on the left; some algebra books write mappings on the right to avoid this.

According to this rule, the product of the two permutations π and σ is the permutation $\pi \circ \sigma = \begin{pmatrix} 1 & 2 & \ldots & n \\ \sigma(\pi(1)) & \sigma(\pi(2)) & \ldots & \sigma(\pi(n)) \end{pmatrix}$.

To multiply the two permutations in tabular form, in the first row of the table corresponding to the permutation π we choose an arbitrary element i. We obtain $\pi(i)$ from the second row of π corresponding to i. Then we find

this number $\pi(i)$ in the first row of the table corresponding to the permutation σ. In the second row of the second table just under this number $\pi(i)$ we find the number $\sigma(\pi(i))$. This is the image of i under the product $\pi \circ \sigma$. We can write it using the following convenient scheme:

$$\begin{array}{cccc} 1 & 2 & \dots & n \\ \downarrow & \downarrow & \dots & \downarrow \\ \pi(1) & \pi(2) & \dots & \pi(n) \\ \downarrow & \downarrow & \dots & \downarrow \\ \sigma(\pi(1)) & \sigma(\pi(2)) & \dots & \sigma(\pi(n)) \end{array} .$$

To illustrate this we give the following example, where the permutation π is written and done first.

Example. Let $\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 1 & 5 & 2 \end{pmatrix}$ and $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 1 \end{pmatrix}$. Then

$$\pi \circ \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 3 & 1 & 5 & 2 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 2 & 1 & 3 \end{pmatrix},$$

but

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

It is clearly the case that $\pi \circ \sigma \neq \sigma \circ \pi$. Hence multiplication of permutations is not a commutative operation.

The identity permutation is written as $\begin{pmatrix} 1 & 2 & 3 & \dots & n \\ 1 & 2 & 3 & \dots & n \end{pmatrix}$. Since a permutation is a bijection, each permutation has an inverse. It is easy to obtain the inverse permutation of a given permutation. All that we need for that is to interchange the upper row with the lower one, and then list the entries in the upper row in ascending order, making the corresponding position change in the bottom row of elements.

Example. Find the inverse of the permutation

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

First flip the upper and lower rows and then rearrange the elements of the upper row in ascending order:

$$\pi^{-1} = \begin{pmatrix} 2 & 3 & 4 & 5 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 1 & 2 & 3 & 4 \end{pmatrix}.$$

It is easy to check that the permutation obtained is the inverse element for π as we see below since

$$\begin{pmatrix}1&2&3&4&5\\2&3&4&5&1\end{pmatrix}\circ\begin{pmatrix}1&2&3&4&5\\5&1&2&3&4\end{pmatrix}=\begin{pmatrix}1&2&3&4&5\\1&2&3&4&5\end{pmatrix},$$

and

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

Definition 1.3.9. *The permutation $\iota=\iota_{kt}$ of the set A is called a transposition (more precisely, the transposition of the symbols $k,t\in A$) if $\iota(k)=t,\iota(t)=k$, and $\iota(j)=j$ for all other elements $j\in A$.*

Thus a transposition interchanges two symbols and fixes the rest of them.

Example. $\begin{pmatrix}1&\mathbf{2}&3&\mathbf{4}&5\\1&\mathbf{4}&3&\mathbf{2}&5\end{pmatrix}$ is a transposition.

We say that the natural numbers m,j *form an inversion pair relative to the permutation* π, if $m<j$ but $\pi(m)>\pi(j)$. For **example**, the permutation $\begin{pmatrix}1&2&3&4\\1&4&3&2\end{pmatrix}$ contains three inversion pairs namely $(2,3),(2,4)$, and $(3,4)$

We let $\mathbf{inv}(\pi)$ denote the number of inversion pairs, relative to the permutation π. We define $\mathbf{sign}\,\pi=(-1)^{\mathbf{inv}(\pi)}$ and call $\mathbf{sign}\,\pi$ *the signature of the permutation* π.

In our last example, $\mathbf{sign}\,\pi=(-1)^3=-1$.

Definition 1.3.10. *The permutation π is called even, if* $\mathbf{sign}\,\pi=1$ *and π is called odd, if* $\mathbf{sign}\,\pi=-1$. *Thus π is even precisely when the number of inversion pairs of π is even and odd when the number of inversion pairs is odd.*

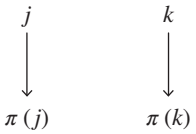
A short computation shows that the equation

$$\mathbf{sign}\,(\pi\circ\sigma)=\mathbf{sign}\,\pi\;\mathbf{sign}\,\sigma$$

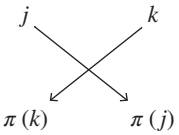
is valid for any permutations π and σ of the same degree. The equation $\mathbf{sign}\,(\pi\circ\sigma)=\mathbf{sign}\,\pi\;\mathbf{sign}\,\sigma$ implies that the product of two even permutations is even, the product of two odd permutations is even, and the product of an even and an odd permutation is odd.

There is a very convenient pictorial method for deciding whether a given permutation π is odd or even, based on the following observation. We rewrite

the permutation π as two rows of numbers, both in the order $1,2,\ldots,n$ and then draw a line from each number k to its image $\pi(k)$ in the second row. Let $1 \leq j < k \leq n$. If (j,k) is not an inversion pair then the two lines drawn from j to $\pi(j)$ and from k to $\pi(k)$ will not intersect. If the lines do intersect then this tells us that (j,k) is an inversion pair and the number of such crossovers for all pairs (j,k) determines the number of these. If numbers j and k don't form an inversion pair relative to π, we obtain a picture of the following type:

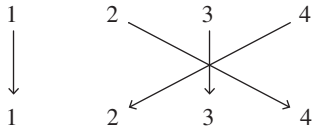


with no crossover of the corresponding lines. If numbers j and k make an inversion pair relative to π, we will have the following picture:



The total number of intersections of these lines is the number of inversion pairs.

We will illustrate this with the following example. Use the permutation we already used above: $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix}$. The following diagram corresponds to this permutation:



As we can see, there are three intersections corresponding to the three pairs of indices forming inversions $(2,3)$, $(2,4)$, and $(3,4)$. In practice we often write the permutation π in the usual fashion, the first row consisting of the elements $\{1,2,\ldots,n\}$ listed in that order. Then we draw lines from each number in the upper row *to the same number* in the bottom row. This is clearly equivalent to the procedure described above.

We let $\mathbf{A}_n$ denote the subset of $\mathbf{S}_n$ consisting of all even permutations. It is not difficult to prove that $|\mathbf{A}_n| = \frac{n!}{2}$.

The representation of permutations as products of cycles plays an important role in their study and we briefly discuss this idea next.

Definition 1.3.11. *Let $1 \leq r \leq n$. A permutation π is called a cycle of length k if there are natural numbers $j_1, \ldots, j_k$ such that*

$$\pi(j_1) = j_2, \pi(j_2) = j_3, \ldots, \pi(j_{k-1}) = j_k, \pi(j_k) = j_1.$$

and $\pi(s) = s$ for all $s \notin \{j_1, \ldots, j_k\}$. The cycle is denoted by $(j_1 j_2 \ldots j_k)$. The numbers $j_1, \ldots, j_k$ are called the elements of this cycle.

In other words, the permutation π "cycles" the indices $j_1, j_2, \ldots, j_k$ around (thus $j_1 \mapsto j_2 \mapsto j_3 \mapsto \ldots \mapsto j_r \mapsto j_1$) but leaves all other indices fixed.

For **example**,

$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 2 & 3 & 7 & 5 & 6 & 4 \end{pmatrix}$ is the cycle (47).

The identity permutation is written as the cycle $(1) = (2) = \ldots$ of length 1. The cycles of length 2 are precisely the transpositions. Notice also that, in this notation, it does not matter which j_r is listed first. Thus permuting the elements of a cycle in cyclic order gives us the same permutation, as for example $(235) = (352) = (523)$.

It is easy to check that cycles with no elements in common (e.g., (13) and (245)) commute with each other and therefore the order of writing the factors is not important in such a case. The following theorem illustrates the importance of cycles.

Theorem 1.3.12. *Every permutation can be represented as a product of cycles with no elements in common and this representation is unique to within the order of the factors.*

We shall not prove this theorem but illustrate the idea of the proof using the following example. Let

$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 3 & 2 & 1 \end{pmatrix}$. We see that π transforms 1 to 4, 4 to 3, 3 to 6, 6 to 1, which means that the cycle (1436) is part of the product decomposition. The permutation π transforms the remaining number 2 to 5, and 5 to 2. Therefore the transposition (25) is also part of the decomposition. So $\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 3 & 2 & 1 \end{pmatrix} = (1436)(25) = (25)(1436)$. The reader can now probably imagine how a general proof would work.

Exercise Set 1.3

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

1.3.1. Let A be a nonempty set. Prove that A is infinite if and only if $\mathbf{S}(A)$ is infinite.

1.3.2. Prove that there is a bijective mapping from $A \times B$ to $B \times A$.

1.3.3. Let A be a set consisting of two elements. Is the multiplication on the set $\mathbf{S}(A)$ commutative?

1.3.4. Let f: $\mathbb{N} \longrightarrow \mathbb{Z}$ be a mapping defined by the rule

$$f(n) = \begin{cases} \dfrac{n}{2} - 1 \text{ whenever } n \text{ is even,} \\ -\dfrac{n+1}{2} \text{ whenever } n \text{ is odd.} \end{cases}$$

Is f injective? If yes, find an inverse to f.

1.3.5. Let $f : \mathbb{Q} \longrightarrow \mathbb{Q}$ be a mapping defined by the rule $f(x) = 3x - |x|$, for $x \in \mathbb{Q}$. Is f injective? If yes, find an inverse for f.

1.3.6. Let $f : \mathbb{Q} \longrightarrow \mathbb{Q}$ be the mapping defined by $f(x) = 2x + |x|$, for $x \in \mathbb{Q}$. Is f injective? If yes, find an inverse for f.

1.3.7. Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be the mapping defined by

$$f(x) = \begin{cases} x^2 \text{ whenever } x \geq 0, \\ x(x-3) \text{ whenever } x < 0. \end{cases}$$

Is f injective? If yes, find an inverse for f.

1.3.8. Let $f : \mathbb{N} \times \mathbb{N} \longrightarrow \mathbb{N}$ be the mapping defined by $f(n,m) = 2^{n-1}(2m-1)$. Is f injective? If yes, find an inverse for f.

1.3.9. Let $f : \mathbb{Q} \longrightarrow \mathbb{Q}$ be the mapping defined by $f(x) = x^2 + 2$, and let $g : \mathbb{Q} \longrightarrow \mathbb{Q}$ be a mapping defined by $g(x) = \frac{x}{2} - 2$. Find the products $g \circ f, f \circ g, (f \circ g) \circ f$, and $f \circ (g \circ f)$

1.3.10. Let $f : \mathbb{Q} \longrightarrow \mathbb{Q}$ be the mapping defined by $f(x) = (1 + (1-x)^{\frac{1}{3}})^{\frac{1}{5}}$. Represent f as a product of four mappings.

1.3.11. Let $f : A \longrightarrow B, g : A \longrightarrow C$. Prove that if f, g are injective then so is $f \circ g$ and that if f, g are surjective then so is $f \circ g$.

1.3.12. Two cycles $(a_1 a_2 a_3 \ldots a_n)$ and $(b_1 b_2 b_3 \ldots b_r)$ are disjoint if $\{a_1, a_2, a_3, \ldots, a_n\} \cap \{b_1, b_2, b_3, \ldots, b_r\} = \emptyset$. Prove that disjoint cycles commute.

1.3.13. Write $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 3 & 6 & 5 & 11 & 7 & 9 & 8 & 1 & 10 & 2 & 4 \end{pmatrix}$ as a product of disjoint cycles and then as a product of transpositions.

1.3.14. Represent the permutation $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 6 & 4 & 5 & 7 & 2 & 8 & 3 & 9 & 1 \end{pmatrix}$ as a product of transpositions and find whether it is even or odd.

1.3.15. Find whether the permutation $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 3 & 1 & 4 & 7 & 5 & 6 \end{pmatrix}$ is even or odd.

1.3.16. Write $(123)(45)(1543)(276)$ first as a product of disjoint cycles, then as a product of transpositions, and then find whether it is even or odd.

1.3.17. Let α be a cycle of length r. Prove that $\alpha^r = \varepsilon$ and that r is the least natural number for which this is true.

1.3.18. If α and β are disjoint cycles of lengths r, s, respectively, then prove that $(\alpha\beta)^l = \varepsilon$, where l is the least common multiple of r and s. Prove also that l is the least natural number for which this is true.

1.3.19. Find the inverse of the permutation $(a_1\, a_2 \ldots a_k)$.

1.3.20. Find $\alpha \circ \beta$ if

$$\alpha = (13)(1468)(26754) \text{ and } \beta = (356)(275)(8941).$$

1.4 OPERATIONS ON MATRICES

In this section we construct some useful examples—matrices—which can be used to illustrate the most important concepts of abstract algebraic structures. Additionally, however, matrices are one of the most useful and prevalent objects in mathematics and its applications. The language of matrices is very convenient and efficient, so is used by scientists everywhere. Matrices are also a central concept in linear algebra, which itself is useful in many fields.

An $m \times n$ matrix is a rectangular table of entries (or elements), containing m rows and n columns which may be numbers or, more generally, any abstract quantities that can be added and multiplied. If the number of rows is equal to the number of columns, then the matrix is called a *square (or quadratic) matrix*, and the number n of its rows (or columns) is called *the order of the matrix*. A matrix of order n is also called an $n \times n$ matrix, the first n refers to the number of rows and the second one to the number of columns. In this book we will mostly consider square matrices. Usually the matrices we consider will have at least order 2.

The choice of the entries used in a matrix depends on the branch of science in which they are used and on the specific problems to be solved. They could be numbers, or polynomials, or functions, or elements of some abstract algebraic structure. In this book we mostly consider numerical matrices, those matrices with numbers as elements.

We denote the entries of a matrix using lower case letters with two indices, which can be thought of as the coordinates of an element in the matrix. The first index shows the number of the row in which the element is situated, while the second index is the number of the place of the element in this row, or, which is the same, the number of the column in which the element lies. Thus an $n \times n$ matrix has the following form:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2,n-1} & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{n,n-1} & a_{nn} \end{pmatrix};$$

square brackets may also be used as in

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2,n-1} & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{n,n-1} & a_{nn} \end{bmatrix}.$$

We call a_{ij} the (i,j) entry of the matrix, so we list the row index first and the column index second. Thus a_{ij} is the entry of the matrix in row i, column j. We also shall use the following brief form for matrix notation

$$[a_{ij}]_{1 \le i,j \le n} \text{ or } [a_{ij}],$$

when the order is reasonably clear.

The set of $n \times n$ matrices whose entries belong to some set S will be denoted by $\mathbf{M}_n(S)$. In this book, we think of S as being a subset of the set, $\mathbb{R}$, of real numbers. In this case, we shall sometimes say that we are dealing with numerical matrices.

We make the following definition of equality of matrices.

Definition 1.4.1. *Two matrices*

$$A = [a_{ij}] \text{ and } B = [b_{ij}]$$

in the set $\mathbf{M}_n(S)$ *are said to be equal, if* $a_{ij} = b_{ij}$ *for every pair of indices* (i,j), *where* $1 \le i,j \le n$.

Thus, equal matrices should have the same order and the same elements in the corresponding places. Certain special types of matrices occur frequently and we next define some of these.

Definition 1.4.2. *Let $A = [a_{ij}]$ be an $n \times n$ numerical matrix.*

(i) *A is called upper triangular, if $a_{ij} = 0$ whenever $i > j$;*
(ii) *If A is upper triangular then A is called unitriangular, if $a_{ii} = 1$ for each $i, 1 \leq i \leq n$;*
(iii) *If A is upper triangular then A is called zero-triangular if $a_{ii} = 0$ for each $i, 1 \leq i \leq n$;*
(iv) *A is called diagonal, if $a_{ij} = 0$ for every $i \neq j$.*

For **example**, the matrix $\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix}$ is *upper triangular*, the matrix $\begin{pmatrix} 1 & a_{12} & a_{13} \\ 0 & 1 & a_{23} \\ 0 & 0 & 1 \end{pmatrix}$ is *unitriangular*, the matrix $\begin{pmatrix} 0 & a_{12} & a_{13} \\ 0 & 0 & a_{23} \\ 0 & 0 & 0 \end{pmatrix}$ is *zero-triangular*, and the matrix $\begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}$ is *diagonal*.

The power of matrices is perhaps best utilized as a means of storing information. An important part of this is concerned with certain natural operations defined on matrices, which we consider next. Just as we can build an arithmetic of numbers so we can build an arithmetic of matrices.

Definition 1.4.3. *Let $A = [a_{ij}]$ and $B = [b_{ij}]$ be matrices in the set $\mathbf{M}_n(\mathbb{R})$. The sum $A + B$ of these matrices is the matrix $C = [c_{ij}] \in \mathbf{M}_n(\mathbb{R})$, whose entries are $c_{ij} = a_{ij} + b_{ij}$ for every pair of indices (i,j), where $1 \leq i,j \leq n$.*

Here is a very easy example to illustrate matrix addition.

Example. $\begin{pmatrix} 1 & 3 \\ 5 & 2 \end{pmatrix} + \begin{pmatrix} 2 & 1 \\ 4 & 3 \end{pmatrix} = \begin{pmatrix} 1+2 & 3+1 \\ 5+4 & 2+3 \end{pmatrix} = \begin{pmatrix} 3 & 4 \\ 9 & 5 \end{pmatrix}$.

The definition means that we can only add matrices if they have the same order and then to add two matrices of the same order we just add the corresponding entries of the two matrices. In this way matrix addition is reduced to the addition of the corresponding entries. Therefore the operation of matrix addition inherits all the properties of number addition.

For example, let $A, B, C \in \mathbf{M}_n(\mathbb{R})$. Then addition of matrices is commutative, which means that $A+B = B+A$. This follows since $a_{ij}+b_{ij} = b_{ij}+a_{ij}$, where $A = [a_{ij}], B = [b_{ij}]$. Likewise, addition of matrices is associative, which means that $(A+B)+C = A+(B+C)$. The set $\mathbf{M}_n(\mathbb{R})$ has a zero matrix O each of whose entries is 0. The matrix O is called the (additive) identity element since $A+O = A = O+A$ for each matrix $A \in \mathbf{M}_n(\mathbb{R})$. It is not hard to see that for each n there is precisely one $n \times n$ matrix with the property that when it is added to A the result is again A. If $A = [a_{ij}]$ then the $n \times n$ matrix $-A$ is the matrix whose entries are $-a_{ij}$. It is easy to see from the definition of matrix addition that $A+(-A) = O = -A+A$. This matrix $-A$ is the unique matrix with the property that when it is added to A the result is the matrix O. The matrix $-A$ is called the additive inverse of the matrix A. Matrix subtraction can be introduced in $\mathbf{M}_n(\mathbb{R})$ by using the natural rule that $A - B = A + (-B)$. This amounts to simply subtracting corresponding entries of the matrix.

Compared to addition, matrix multiplication looks more sophisticated, and does not seem as natural as addition.

Definition 1.4.4. *Let $A = [a_{ij}]$ and $B = [b_{ij}]$ be two matrices in the set $\mathbf{M}_n(\mathbb{R})$. The product AB of these matrices is the matrix $C = [c_{ij}]$, whose elements are*

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^{n} a_{ik}b_{kj}$$

for every pair of indices (i,j), where $1 \le i,j \le n$.

Thus to obtain the (i,j) entry of the product C, we need to multiply pairwise the elements of row i of the matrix A by the corresponding elements of column j of the matrix B and add the results.

Example. Let $A = \begin{pmatrix} 1 & 3 \\ 5 & 2 \end{pmatrix}, B = \begin{pmatrix} 2 & 1 \\ 4 & 3 \end{pmatrix}$,

$$\begin{pmatrix} \mathbf{1} & \mathbf{3} \\ 5 & 2 \end{pmatrix} \begin{pmatrix} \mathbf{2} & 1 \\ \mathbf{4} & 3 \end{pmatrix} = \begin{pmatrix} 1\times 2+3\times 4 = 14 & \\ & \end{pmatrix},$$

$$\begin{pmatrix} 1 & 3 \\ \mathbf{5} & \mathbf{2} \end{pmatrix} \begin{pmatrix} \mathbf{2} & 1 \\ \mathbf{4} & 3 \end{pmatrix} = \begin{pmatrix} 1\times 2+3\times 4 = 14 & \\ 5\times 2+2\times 4 = 18 & \end{pmatrix},$$

$$\begin{pmatrix} \mathbf{1} & \mathbf{3} \\ 5 & 2 \end{pmatrix} \begin{pmatrix} 2 & \mathbf{1} \\ 4 & \mathbf{3} \end{pmatrix} = \begin{pmatrix} 1\times 2+3\times 4 = 14 & 1\times 1+3\times 3 = 10 \\ 5\times 2+2\times 4 = 18 & \end{pmatrix},$$

$$\begin{pmatrix} 1 & 3 \\ \mathbf{5} & \mathbf{2} \end{pmatrix} \begin{pmatrix} 2 & \mathbf{1} \\ 4 & \mathbf{3} \end{pmatrix} = \begin{pmatrix} 1\times 2+3\times 4 = 14 & 1\times 1+3\times 3 = 10 \\ 5\times 2+2\times 4 = 18 & 5\times 1+2\times 3 = 11 \end{pmatrix}.$$

So, $\begin{pmatrix} 1 & 3 \\ 5 & 2 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 4 & 3 \end{pmatrix} = \begin{pmatrix} 14 & 10 \\ 18 & 11 \end{pmatrix}$.

Matrix multiplication is not commutative as the previous example shows since

$$\begin{pmatrix} 2 & 1 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} 1 & 3 \\ 5 & 2 \end{pmatrix} = \begin{pmatrix} 7 & 8 \\ 19 & 18 \end{pmatrix} \neq \begin{pmatrix} 1 & 3 \\ 5 & 2 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 4 & 3 \end{pmatrix}.$$

The matrices A, B satisfying $AB = BA$ are called permutable (in other words these matrices commute), and not all matrices have this property. However, matrix multiplication does possess other important properties, namely, the associative and distributive properties, which are exhibited in the following theorem.

Theorem 1.4.5. *For arbitrary matrices $A, B, C \in \mathbf{M}_n(\mathbb{R})$, the following properties hold:*

(i) $(AB)C = A(BC)$;
(ii) $(A+B)C = AC+BC$;
(iii) $A(B+C) = AB+AC$;
(iv) *There exists a matrix $I = I_n \in \mathbf{M}_n(\mathbb{R})$ such that $AI = IA = A$ for each matrix $A \in \mathbf{M}_n(\mathbb{R})$. For a given value of n, I is the unique matrix with this property.*

Proof.

(i) By definition, we need to show that the corresponding entries of $(AB)C$ and $A(BC)$ are equal. To do this, let

$$A = [a_{ij}], B = [b_{ij}], \text{ and } C = [c_{ij}].$$

Put

$$AB = [d_{ij}], BC = [v_{ij}],$$
$$(AB)C = [u_{ij}], A(BC) = [w_{ij}].$$

We must show that $u_{ij} = w_{ij}$ for arbitrary (i,j), where $1 \leq i,j \leq n$. We have

$$u_{ij} = \sum_{1\leq k\leq n} d_{ik}c_{kj} = \sum_{1\leq k\leq n}\left(\sum_{1\leq m\leq n} a_{im}b_{mk}\right)c_{kj} = \sum_{1\leq k\leq n}\sum_{1\leq m\leq n}(a_{im}b_{mk})c_{kj}$$

and

$$w_{ij} = \sum_{1\le m\le n} a_{im}v_{mj} = \sum_{1\le m\le n} a_{im}\left(\sum_{1\le k\le n} b_{mk}c_{kj}\right)$$
$$= \sum_{1\le m\le n}\sum_{1\le k\le n} a_{im}(b_{mk}c_{kj}) = \sum_{1\le k\le n}\sum_{1\le m\le n} a_{im}(b_{mk}c_{kj}).$$

Since $(a_{im}b_{mk})c_{kj} = a_{im}(b_{mk}c_{kj})$ it follows that $u_{ij} = w_{ij}$ for all pairs (i,j). Hence $(AB)C = A(BC)$.

(ii) We need to show that corresponding entries of $AC+BC$ and $A(B+C)$ are equal. Put

$$AC = [x_{ij}], BC = [y_{ij}], (A+B)C = [z_{ij}].$$

We shall prove that $z_{ij} = x_{ij} + y_{ij}$ for arbitrary i,j, where $1 \le i,j \le n$. We have

$$z_{ij} = \sum_{1\le k\le n} (a_{ik}+b_{ik})c_{kj} = \sum_{1\le k\le n} a_{ik}c_{kj} + \sum_{1\le k\le n} b_{ik}c_{kj} = x_{ij} + y_{ij}.$$

Thus $(A+B)C = AC+BC$.

The proof of (iii) is similar.

(iv) We define the symbol δ_{ij} (the Kronecker delta) by

$$\delta_{ij} = \begin{cases} 0, \text{ if } i \ne j, \\ 1, \text{ if } i = j. \end{cases}$$

It is easy to check that

$$I = [\delta_{ij}] = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

has the required property that $AI = IA = A$.

In order to prove the uniqueness of I assume that there also exists a matrix U such that $AU = UA = A$ for each matrix $A \in \mathbf{M}_n(\mathbb{R})$. Setting $A = I$ we obtain $IU = I$. Also, though, we know that $IU = U$, from the definition of I, so that $I = U$.

The matrix $I = I_n$ is called the $n \times n$ *identity matrix*.

Definition 1.4.6. *Let $A \in \mathbf{M}_n(\mathbb{R})$. The matrix $B \in \mathbf{M}_n(\mathbb{R})$ is called an inverse (or reciprocal) of A if $AB = BA = I$. The matrix A is then said to be invertible or non-singular.*

Many nonzero matrices lack inverses. For example, consider the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Then for an arbitrary matrix $\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$ we have $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ 0 & 0 \end{pmatrix}$. This product would never be the identity matrix, thus our matrix has no inverse.

If a matrix A has an inverse, then this inverse matrix is unique. To see this let U, V be inverses of the matrix A so that $AU = UA = I = VA = AV$. Then we have

$$V(AU) = VI = V \text{ and } V(AU) = (VA)U = IU = U.$$

Thus $V = U$. We follow the usual convention and denote the inverse of the matrix A by A^{-1}.

We note that criteria for the existence of an inverse for a given matrix are closely connected to the idea of the determinant of a matrix, a concept usually introduced in a linear algebra course, where the properties of determinants are investigated. We shall need only one result, which states that a matrix $A \in \mathbf{M}_n(\mathbb{R})$ has a multiplicative inverse if and only if its determinant, $\mathbf{det}(A)$, is nonzero. The matrix A is called nonsingular in this case and singular if $\mathbf{det}(A) = 0$. The proof can be found in any treatise on linear algebra.

Now we consider multiplication of a matrix by a number, or scalar.

Definition 1.4.7. *Let $A = [a_{ij}]$ be a matrix from the set $\mathbf{M}_n(\mathbb{R})$ and let $\alpha \in \mathbb{R}$. The product of α and the matrix A is the matrix $\alpha A = [c_{ij}] \in \mathbf{M}_n(\mathbb{R})$, whose entries are defined by $c_{ij} = \alpha a_{ij}$, for every pair of indices (i,j), where $1 \le i, j \le n$.*

Thus, when we multiply a matrix by a real number we multiply each element of the matrix by this number. Here are the main properties of this operation, which can be proved quite easily, in a manner similar to that given in Theorem 1.4.5. We note that these equations hold for all real numbers α, β and for all matrices A, B where the multiplication is defined.

Theorem 1.4.8. *Let A, B be matrices and α, β real numbers.*

(i) $(\alpha + \beta)A = \alpha A + \beta A$;
(ii) $\alpha(A + B) = \alpha A + \alpha B$;
(iii) $\alpha(\beta A) = (\alpha\beta)A$;

(iv) $1A = A$;

(v) $\alpha(AB) = (\alpha A)B = A(\alpha B)$.

Note that this operation of multiplying a matrix by a number can be reduced to the multiplication of two matrices since $\alpha A = (\alpha I)A$.

Here is a summary of all the properties we have obtained so far, using our previously established notation.

$A+B = B+A$,
$A+(B+C) = (A+B)+C$,
$A+O = A$,
$A+(-A) = O$,
$A(B+C) = AB+AC$,
$(A+B)C = AC+BC$,
$A(BC) = (AB)C$,
$AI = IA = A$,
$(\alpha+\beta)A = \alpha A+\beta A$,
$\alpha(A+B) = \alpha A+\alpha B$,
$\alpha(\beta A) = (\alpha\beta)A$,
$1A = A$,
$\alpha(AB) = (\alpha A)B = A(\alpha B)$.

Exercise Set 1.4

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

1.4.1. Let A be a diagonal matrix. Suppose all entries on the main diagonal are different. Let B be a matrix such that $AB = BA$. Prove that B is diagonal.

1.4.2. Find all matrices $A \in M_2(\mathbb{R})$ with the property $A^2 = O$.

1.4.3. Let A and B be matrices. If we interchange the m-th and t-th rows of A, what changes does this imply in the matrix AB?

1.4.4. Let $A, B \in \mathbf{M}_n(\mathbb{R})$. If $\alpha \in \mathbb{R}$ and if we add α times row t to row m in the matrix A then what changes does this imply in the matrix AB?

1.4.5. Find A^3 if $A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$.

1.4.6. Find $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}^3$.

1.4.7. Find $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}^3$.

1.4.8. Find $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 0 & 1 & 0 \\ 4 & 0 & 0 & 1 \end{pmatrix}^3$.

1.4.9. Find $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}^4$.

1.4.10. Let $A \in \mathbf{M}_n(\mathbb{R})$. A matrix A is called nilpotent, if $A^k = O$ for some positive integer k. The minimal such number k is called the nilpotency class of A. Prove that every zero triangular matrix is nilpotent.

1.4.11. Let $A \in \mathbf{M}_2(\mathbb{R})$ be a matrix such that $AX = XA$ for all $X \in \mathbf{M}_2(\mathbb{R})$. Prove that $A = rI$ for some $r \in \mathbb{R}$. What will the general form of this result be?

1.4.12. If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and if $ad - bc \neq 0$ then show that $A^{-1} = \dfrac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

1.4.13. Solve the following matrix equation $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} X = \begin{pmatrix} 3 & 5 \\ 5 & 9 \end{pmatrix}$.

1.4.14. Find the matrix products: $\begin{pmatrix} 3 & -2 \\ 4 & -1 \end{pmatrix}\begin{pmatrix} -5 & 2 \\ 2 & 4 \end{pmatrix}$ and $\begin{pmatrix} 1 & -1 & 4 \\ 3 & -2 & 7 \\ -2 & 3 & -4 \end{pmatrix}\begin{pmatrix} -2 & 0 & 3 \\ 0 & -4 & 7 \\ 2 & -1 & 5 \end{pmatrix}$.

1.4.15. Prove that if A, B are invertible matrices of $\mathbf{M}_n(\mathbb{R})$ then AB is invertible and find a formula for its inverse in terms of A^{-1} and B^{-1}.

1.4.16. Prove that if $A \in \mathbf{M}_n(\mathbb{R})$ is such that $A^n = O$ then $I + A$ is invertible.

1.4.17. Find all matrices $A \in \mathbf{M}_2(\mathbb{R})$ such that $A^2 = I$.

1.4.18. If $A = [a_{ij}] \in \mathbf{M}_n(\mathbb{R})$ then the transpose of A is the matrix $A^t = [b_{ij}]$ where $b_{ij} = a_{ji}$. Show that $(AB)^t = B^t A^t$ whenever also $B \in \mathbf{M}_n(\mathbb{R})$.

1.4.19. A matrix $A = [a_{ij}] \in \mathbf{M}_n(\mathbb{R})$ is symmetric if $a_{ji} = a_{ij}$ for all $i \neq j$ and skew symmetric if $a_{ji} = -a_{ij}$ when $i \neq j$. Prove the following facts: (a) $A + A^t$ is symmetric, (b) $A - A^t$ is skew symmetric.

1.4.20. Prove that every square matrix is a sum of a symmetric matrix and a skew symmetric matrix, in a unique way.

1.5 BINARY ALGEBRAIC OPERATIONS AND EQUIVALENCE RELATIONS

In this section we are interested in binary (algebraic) operations; these are important in mathematics and certainly in modern algebra. Indeed, modern algebra could be regarded as a branch of mathematics that studies algebraic operations, since much of the time we are not interested in the nature of the elements of a set, but are more interested in how an algebraic operation defined on the set acts on the elements of the set.

The usual addition and multiplication of two rational numbers are just two examples of binary operations defined on the set $\mathbb{Q}$ of rational numbers. The main idea here is that associated with every ordered pair of rational numbers there is another rational, its sum or product. Thus addition and multiplication can really be thought of as mappings of the Cartesian product $\mathbb{Q} \times \mathbb{Q}$ to $\mathbb{Q}$.

As another example of this concept we recall that in Section 1.3 the set of permutations $\mathbf{S}_n$ was introduced for each natural number n. There we saw how to define the product of two permutations in $\mathbf{S}_n$, which we now view as a mapping from the Cartesian product $\mathbf{S}_n \times \mathbf{S}_n$ to $\mathbf{S}_n$ and this is also a binary operation on the set of such permutations. In this case, we recall that the multiplication is not commutative.

Definition 1.5.1. *Let M be a set. The mapping $\theta : M \times M \longrightarrow M$ from the Cartesian square of M to M is called a binary (algebraic) operation on the set M. Thus, corresponding to every ordered pair (a,b) of elements, where $a, b \in M$, there is a uniquely defined element $\theta(a,b) \in M$. The element $\theta(a,b) \in M$ is called the composition of the elements a and b relative to this operation.*

Notice that there are two important ideas here. One is that $\theta(a,b)$ is an element of M; the other is that $\theta(a,b)$ is uniquely determined by the ordered pair (a,b). Furthermore, it is often rather cumbersome to keep referring to the

function θ and using the notation $\theta(a,b)$. Therefore $\theta(a,b)$ is often written $a*b$ or as $a\circ b$ so that $*$ (or $\circ$, or some other symbol) denotes the operation. In many natural cases where addition is involved the operation will usually be denoted by + and the corresponding composition $a+b$ is then called the *sum* of a and b. In this case, we talk about the *additive notation* of the binary operation. Often also, multiplication is the operation involved and in this case the sign $\cdot$ is usually used for multiplication; the corresponding composition $a\cdot b$ is called the *product* of a and b. In this case, we say that *multiplicative notation* is being used. Traditionally, the $\cdot$ is omitted and we will also often follow this convention so that $a\cdot b$ will usually be writen as ab.

We now give some further examples of binary operations. Usually it is quite easy to verify that these are binary operations, but they serve to illustrate that binary operations are very familiar to the reader, as we observed above. The question to be resolved, for a given binary operation $*$ defined on a set M, is whether or not $a*b\in M$, for all $a,b\in M$.

(i) Addition on the sets $\mathbb{N},\mathbb{Z},\mathbb{Q},\mathbb{R}$;

(ii) Multiplication on the sets $\mathbb{N},\mathbb{Z},\mathbb{Q},\mathbb{R}$;

(iii) Let M be a set and let $\mathbf{P}(M)$ be the set of all transformations of M. Then, for all $f,g\in\mathbf{P}(M)$, the map θ defined by $\theta(f,g)=f\circ g$ is a binary operation on the set $\mathbf{P}(M)$;

(iv) Addition and multiplication of matrices in $\mathbf{M}_n(\mathbb{R})$.

(v) Addition and multiplication of real functions (i.e., transformations of the set $\mathbb{R}$).

(vi) Addition of vectors and vector product on the space $\mathbb{R}^3$.

(vii) The mappings$(n,k)\longmapsto n^k,(n,k)\longmapsto n^k+k^n,n,k\in\mathbb{N}$ define binary operations on $\mathbb{N}$.

(viii) The mappings $(n,k)\longmapsto\mathbf{GCD}(n,k)$, the greatest common divisor of n and k, and $(n,k)\longmapsto\mathbf{LCM}(n,k)$, the least common multiple of n and k, define binary operations on $\mathbb{Z}$. (Here the nonnegative $\mathbf{GCD}(n,k)$ and $\mathbf{LCM}(n,k)$ are chosen.)

We now consider some important properties of binary algebraic operations. To be concrete we may use the multiplicative form of writing a binary operation but may also illustrate the additive form. However we stress that our binary operations are very much more general than ordinary addition or multiplication.

Definition 1.5.2. *Let M be a set with a binary algebraic operation $*$. A subset S of M is called* closed *or* stable *with respect to $*$ if for each pair of elements $a,b\in S$ the element $a*b$ also belongs to S.*

This means that the restriction of the binary operation $*$ to S is a binary operation on S.

Example. The set $\mathbb{N}$ is a closed subset of $\mathbb{Z}$ with respect to addition, but not with respect to subtraction since, for example, $1, 2$ are natural numbers but $1-2$ is not. The subset of all even integers is a closed subset of $\mathbb{Z}$ with respect to addition and multiplication, while the subset of odd integers is closed with respect to multiplication, but not addition.

Definition 1.5.3. *A binary operation on a set M is called commutative if $ab = ba$ for each pair a, b of elements of M.*

For the additive form, commutativity of a and b would be written as follows:

$$a+b = b+a, \text{ where } a, b \in M.$$

Concrete examples of commutative operations include: multiplication and addition on the set of integers, rational, and real numbers; matrix addition; multiplication of real functions; the operations **GCD**, **LCM** on $\mathbb{Z}$ and vector addition in $\mathbb{R}^3$. As we saw earlier, multiplication of transformations is not commutative in general. Likewise, multiplication of matrices is not commutative in general.

If we have three elements $a, b, c \in M$, then we can form the products $a(bc)$ and $(ab)c$ and in general these may be different as when we form $(a-b)-c$ and $a-(b-c)$, for real numbers a, b, c.

Definition 1.5.4. *A binary operation on a set M is called associative if $(ab)c = a(bc)$ for all elements a, b, c of M.*

Written additively this becomes

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

The examples mentioned, except for the examples involving the operations

$$(n,k) \longmapsto n^k, (n,k) \longmapsto n^k + k^n,$$

on $\mathbb{N}$, and the vector product on $\mathbb{R}^3$, are associative. Thus, when $n * k = n^k$ then $(2*1)*3 = 8$ whereas $2*(1*3) = 2$.

For four elements a, b, c, d, we can construct a number of different products. For example, we can determine each of the products

$$((ab)c)d, (ab)(cd), (a(bc))d, a(b(cd)) \text{ and } a((bc)d)$$

to name but a few. When the operation is associative, however, all methods of bracketing give the same expression so that there is no need for any complicated bracketing. For example, we have

$(a(bc))d = ((ab)c)d);$
$(ab)(cd) = ((ab)c)d);$
$a(b(cd)) = (ab)(cd) = ((ab)c)d);$
$a((bc)d) = (a(bc))d = ((ab)c)d.$

It can be shown in general that when a binary operation is associative the way in which we position the brackets in an expression makes no difference, assuming that the order of the elements is unchanged. In particular, we do not even need to put brackets in a product of elements $a_1, a_2, \ldots, a_n$ and just write the product as $a_1 a_2 \ldots a_n$ or, more succinctly, $\prod_{1 \le i \le n} a_i$. When needed we can place parentheses in any manner. In the case when $a_1 = a_2 = \cdots = a_n = a$, we will denote the product $a_1 a_2 \ldots a_n$ by a^n, as usual, and call it the n-th power of a.

For an associative binary operation on a set M the usual "rule of exponents" holds, at least for exponents that are natural numbers. Thus for each element $a \in M$ and arbitrary $n, m \in \mathbb{N}$ we have

$$a^n a^m = a^{n+m}, (a^n)^m = a^{nm}.$$

When additive notation is used, instead of multiplicative, powers become multiples; thus instead of $\prod_{1 \le i \le n} a_i$ we write $\sum_{1 \le i \le n} a_i$ and if $a_1 = a_2 = \cdots = a_n$, then write $a_1 + a_2 + \cdots + a_n = na$. In this case the rules of exponents become properties of multiples as follows:

$$na + ma = (n+m)a,\ m(na) = (mn)a.$$

Two elements a, b are said to *commute* or *permute* if $ab = ba$. We also sometimes say a and b are permutable. If the elements a, b commute then $(ab)^n = a^n b^n$ for each $n \in \mathbb{N}$. More generally, if $a_1, a_2, \ldots, a_n$ are elements of M and if the operation on M is commutative and associative, then

$$(a_1 a_2 \ldots a_n)^m = a_1^m a_2^m \ldots a_n^m$$

for every $m \in \mathbb{N}$. Additively this would be written as

$$m(a_1 + a_2 + \cdots + a_n) = ma_1 + ma_2 + \cdots + ma_n.$$

Definition 1.5.5. *Let M be a set with binary operation $*$. The element $e \in M$ is called a neutral (or identity) element under this operation if $a * e = e * a = a$ for each element a of the set M.*

The neutral element of a set M is unique whenever it exists. Indeed, if e_1 is another element with the property $a * e_1 = e_1 * a = a$ for all $a \in M$, then we may let $a = e$ or $a = e_1$, in the definitions of e_1 and e respectively, and then we obtain $e = e_1 * e = e_1$. Sometimes, to avoid ambiguity, we may use the notation e_M for the identity element of M.

If multiplicative notation is used then we use the term *identity element*, and often use the notation 1, or 1_M, for the neutral element e. In this case we have $1 \cdot a = a \cdot 1 = a$, for all $a \in M$. We emphasize that 1_M need not be the integer 1 here. If additive notation is used, then the neutral element is usually called the *zero element* and is often denoted by 0, or 0_M, so that the definition of the zero element is $a + 0 = 0 + a = a$ for each element $a \in M$; again 0_M should not be confused with the integer 0.

Examples.

(i) The operation of addition on the sets of all natural, integer, rational, and real numbers has a zero element, the number 0.

(ii) The operation of multiplication on the sets of all natural, integer, rational, and real numbers has an identity element,the number 1.

(iii) Let M be a set and $\mathbf{P}(M)$ be the set of all transformations of the set M. When the operation is composition of transformations of the set M, the identity element is the permutation $\varepsilon_M : M \longrightarrow M$, defined by $\varepsilon_M(m) = m$, for all $m \in M$.

(iv) The zero matrix is the zero element for the operation of addition on the set $\mathbf{M}_n(\mathbb{R})$ of real matrices whereas the identity matrix I is the identity element when the operation is multiplication on the set $\mathbf{M}_n(\mathbb{R})$.

(v) The function with value 0 for all elements in the domain is the zero element when real functions are added: and when the operation is multiplication the identity element is the function $f(x)$ for which $f(x) = 1$ for all $x \in \mathbb{R}$.

(vi) The operation $n * k = \mathbf{GCD}(n, k)$, whenever $n, k \in \mathbb{Z}$, has neutral element the number 0, since $\mathbf{GCD}(n, 0) = n$, for all $n \in \mathbb{Z}$.

(vii) For addition of vectors in $\mathbb{R}^3$, the zero element is the zero vector.

Definition 1.5.6. *Let M be a set with a binary operation and suppose that there is an identity element e. The element $x \in M$ is called an inverse of the element $a \in M$ if*

$$ax = xa = e.$$

If a has an inverse then we say that a is invertible.

When we use additive notation we will often also use the term "additive inverse" and when we use multiplicative notation we will also use the term "multiplicative inverse."

If the operation on M is associative and the element $a \in M$ is invertible then a has a unique inverse. To see this, let y be an element of M that also satisfies

$$ay = ya = e.$$

Then

$$y = ey = (xa)y = x(ay) = xe = x.$$

We denote the unique inverse of a by a^{-1}. We note that $aa^{-1} = a^{-1}a = e$ and so, evidently,

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

If the operation on M is written additively then we denote the inverse of a, should it exist, by $-a$, called the negative (or sometimes the opposite) of a. In this case the definition of the additive inverse takes the form:

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

Proposition 1.5.7. *Let M be a set with an associative binary operation and suppose that M has an identity element e. If the elements $a_1, a_2, \ldots, a_n$ are invertible in M, then the product $a_1 a_2 \ldots a_n$ is also invertible and*

$$(a_1 a_2 \ldots a_n)^{-1} = a_n^{-1} a_{n-1}^{-1} \ldots a_1^{-1}.$$

Proof. We have, informally,

$$\begin{aligned}(a_1 a_2 \ldots a_n)(a_n^{-1} a_{n-1}^{-1} \ldots a_1^{-1}) &= (a_1 \ldots a_{n-1})(a_n a_n^{-1})(a_{n-1}^{-1} \ldots a_1^{-1}) \\ &= (a_1 a_2 \ldots a_{n-1})(a_{n-1}^{-1} \ldots a_1^{-1}) = \cdots = e.\end{aligned}$$

This shows that the proposition holds, since we have exhibited an element which multiplies $a_1 \ldots a_n$ to give e.

The existence of an identity element and the inverse of an element a allows us to define all integer powers of a. To do this we define

$$a^0 = e, \text{ and } a^{-n} = \left(a^{-1}\right)^n, \text{ whenever } n \in \mathbb{N}.$$

In additive notation these definitions take the form:

$$0a = 0_M \text{ and } (-n)a = n(-a).$$

Our next result shows that the usual rules of exponents hold for all integer powers.

Proposition 1.5.8. *Let M be a set together with an associative binary operation and suppose that M has an identity element e. If $a \in M$ is invertible and $m, n \in \mathbb{Z}$ then*

$$a^n a^m = a^{n+m} \text{ and } (a^n)^m = a^{nm}.$$

Proof. If $n, m > 0$, then the assertion follows by simply writing out the products. Furthermore if one of m or n is 0 then the equalities hold in any case. If $m, n < 0$, then $n = -p, m = -q$, for certain $p, q \in \mathbb{N}$. Then, using the definitions we have,

$$a^n a^m = a^{-p}a^{-q} = (a^{-1})^p (a^{-1})^q = (a^{-1})^{p+q} = a^{-(p+q)} = a^{-p-q} = a^{n+m}$$

and

$$(a^n)^m = (a^{-p})^{-q} = \left((a^p)^{-1}\right)^{-q} = \left(\left((a^p)^{-1}\right)^{-1}\right)^q = (a^p)^q = a^{pq} = a^{mn}.$$

Suppose now that $n > 0$, $-q = m < 0$ and $n > -m = q$. Then

$$a^n a^m = \underbrace{a \dots a}_{n}\underbrace{(a^{-1}) \dots (a^{-1})}_{q} = \underbrace{a \dots a}_{n-q} = a^{n+m}.$$

If $n > 0, -q = m < 0$ and $n < -m = q$, then

$$a^n a^m = \underbrace{a \dots a}_{n}\underbrace{(a^{-1}) \dots (a^{-1})}_{q} = \underbrace{a^{-1} \dots a}_{q-n}^{-1} = (a^{-1})^{-(n+m)} = a^{n+m}.$$

For the second equation, if $n > 0$ and $-q = m < 0$ then

$$(a^n)^m = ((a^n)^{-1})^q = ((a^{-1})^n)^q = (a^{-1})^{nq} = (a^{-1})^{-nm} = a^{-(-nm)} = a^{nm}.$$

If $-p = n < 0, m > 0$, then

$$(a^n)^m = ((a^{-1})^p)^m = (a^{-1})^{pm} = (a^{-1})^{-nm} = a^{-(-nm)} = a^{nm}.$$

The result follows.

Equivalence relations

In Section 1.2 we defined a binary relation on a set A to be a subset of the Cartesian product $A \times A$. If Φ is such a binary relation and if $(x,y) \in \Phi$, then we say that elements x and y (in this given order) correspond to each other via Φ. Instead of the notation $(x,y) \in \Phi$ we will use the notation $x\Phi y$, which is more suggestive, since typically we think of x and y being related by Φ. This form of notation is called *infix*.

Here are the most important properties of binary relations.

Definition 1.5.9. *Let A be a set with a binary relation Φ.*

(i) *Φ is called reflexive if $(a,a) \in \Phi$ (or $a\,\Phi\,a$), for each $a \in A$;*

(ii) *Φ is called transitive if, whenever $a,b,c \in A$ and $(a,b),(b,c) \in \Phi$, then $(a,c) \in \Phi$ (or, alternatively, $a\,\Phi\,b$ and $b\,\Phi\,c$ imply that $a\,\Phi\,c$);*

(iii) *Φ is called symmetric if, whenever $a,b \in A$ and $(a,b) \in \Phi$, then $(b,a) \in \Phi$ (or, alternatively, $a\,\Phi\,b$ implies $b\,\Phi\,a$);*

(iv) *Φ is called antisymmetric if, whenever $a,b \in A$ and $(a,b),(b,a) \in \Phi$ then $a = b$ (or, alternatively, $a\,\Phi\,b$ and $b\,\Phi\,a$ imply $a = b$).*

If A is a finite set we make a pair of perpendicular axes and label the axes with points representing the elements of A. If $a,b \in A$ and $a\Phi b$, then we can plot the point (a,b), as we do in the usual rectangular coordinate system, by finding the point on the horizontal axis labelled a and the point on the vertical axis labelled b and putting a mark (cross or circle) at the place where the lines drawn from these points would intersect. In this way, relations can be pictured.

There are many **examples** of reflexive relations and here we give only a few: If A is the set of all straight lines in the plane then the relation of "being parallel" is certainly reflexive; the relation "looks alike" on a certain set of people is clearly reflexive since everyone looks alike themselves; the relation of "having the same gender" on a set of animals is certainly reflexive and so on.

The relation "x is the brother of y" is symmetric on the set of all males, but is not symmetric on the set of all people, since y will only be the brother of x if y is a male. Here are some examples of transitive relations: the relation "to be divisible by" on the set of integers, the relation "to be greater" on the set of real numbers, the relation "to be older" on a set of people, the relation "to have the same color" on the set of toys, and so on.

The following concept leads us to an important type of binary relation called an equivalence relation.

Definition 1.5.10. *A family $\mathfrak{S}$ of subsets of a set A is called a covering if $A = \cup\mathfrak{S}$ (thus for each $x \in A$ there exists $S \in \mathfrak{S}$ such that $x \in S$). A covering $\mathfrak{S}$ is called a partition of the set A if, additionally, $X \cap Y = \emptyset$, whenever $X, Y \in \mathfrak{S}$ and $X \neq Y$; thus all pairs of distinct subsets of the partition have empty intersection.*

Let $\mathfrak{S}$ be a partition of the set A and define a binary relation $\Gamma(\mathfrak{S})$ on A by the rule that $(x, y) \in \Gamma(\mathfrak{S})$ if and only if the elements x and y belongs to the same set S from the family $\mathfrak{S}$. The relation $\Gamma(\mathfrak{S})$ has various properties which we now discuss. Since $A = \cup\mathfrak{S}$ then, for each element $x \in A$, there exists a subset $S \in \mathfrak{S}$ such that $x \in S$. Thus $(x, x) \in \Gamma(\mathfrak{S})$ and hence the relation $\Gamma(\mathfrak{S})$ is reflexive. It is clear that the relation $\Gamma(\mathfrak{S})$ is symmetric. Finally, let $(x, y), (y, z) \in \Gamma(\mathfrak{S})$. It follows that there exist subsets $S, R \in \mathfrak{S}$ such that $x, y \in S$ and $y, z \in R$. In particular, $y \in S \cap R$ and using the definition of a partition, we see that $S = R$. Hence the elements x, z belongs to S which is an element of the partition $\mathfrak{S}$. Thus $(x, z) \in \Gamma(\mathfrak{S})$ so the relation $\Gamma(\mathfrak{S})$ is transitive.

Definition 1.5.11. *A binary relation Φ on a set A is called an equivalence relation or an equivalence if it is reflexive, symmetric, and transitive.*

We give some examples next. First we say that two polygons are equivalent if they have the same number of vertices. Thus, for example, under this relation all triangles are equivalent, and it is easy to see that this relation is an equivalence relation. The family of all triangles can itself be partitioned into the subsets of acute, right-angled, and obtuse triangles and this partition helps define an equivalence relation on the set of all triangles. We can also say that two triangles are equivalent depending upon whether they are scalene, isosceles, or equilateral. Thus a given set may have more than one equivalence relation defined on it. More generally, the relation "the figure A is similar to the figure B" on the set of all geometric figures is an equivalence relation. We note too that our work here shows that every partition $\mathfrak{S}$ of a set A gives rise to an equivalence relation $\Gamma(\mathfrak{S})$ defined on A.

One main reason for studying equivalence relations is that such relations allow us to construct new mathematical objects quite rigorously. For example, the relation of colinearity of rays is a partition of the plane or space into classes of colinear rays. Each of these classes is called a direction, or a path. In this way we can transform the intuitive idea of direction into a rigorously defined concept. In a similar way, given a collection of figures we can define a relation on this set of figures by saying that figure A is related to figure B if and only if A has the same shape as B. Children forever use partitions (and hence equivalence relations!) in their play. For example a child might sort its toys according to color and the relation "is the same color as" is an equivalence relation.

Here is a list of some further **examples** of equivalence relations.

(i) If A is an arbitrary set there are two extreme cases: $\Phi = A \times A$ and $\Phi = \{(x,x) \mid x \in A\}$ (the diagonal of the Cartesian product $A \times A$). These are both examples of equivalence relations and all other equivalence relations on A are situated between these two extreme cases;
(ii) the relation "to be parallel" on the set of all straight lines in a plane;
(iii) the relation of similarity;
(iv) the relation "to be equivalent equations" on the set of equations;
(v) the relation "to belong to the same species" on the set of animals;
(vi) the relation "to be relatives" on the set of people;
(vii) the relation "to be the same height" on the set of people;
(viii) the relation "to live in the same city" on the set of people;
(ix) the relation "has the same birthday as" on the set of all people;
(x) the relation "is similar to" or "congruent to" on the set of all triangles;
(xi) the relation "has the same image" on the elements of the domain of a function.

We have already seen that each partition of a set generates an equivalence relation. We now show that, conversely, each equivalence relation on a set leads to a natural partition of the set.

Definition 1.5.12. *Let Φ be an equivalence relation on the set A and let $x \in A$. The subset $[x]_\Phi = \{y \in A \mid (x,y) \in \Phi\}$ is called the equivalence class of x.*

Thus the equivalence class of x consists precisely of those elements of A that are equivalent to x. It is important to note that each equivalence class is uniquely defined by each of its elements. Indeed, let $y \in [x]_\Phi$ so that $(x,y) \in \Phi$. If $z \in [y]_\Phi$, then $(y,z) \in \Phi$ also. Since the equivalence relation is transitive it follows that $(x,z) \in \Phi$ also and hence $z \in [x]_\Phi$. Thus $[y]_\Phi \subseteq [x]_\Phi$. Because equivalence relations are symmetric we also have $[x]_\Phi \subseteq [y]_\Phi$ and hence $[x]_\Phi = [y]_\Phi$.

Since $(x,x) \in \Phi$, it follows that $x \in [x]_\Phi$ and hence the family of all equivalence classes forms a covering set of A. Next we consider the intersection, $[x]_\Phi \cap [y]_\Phi$, of two equivalence classes and suppose that this intersection is not empty. Let $z \in [x]_\Phi \cap [y]_\Phi$. Then, as we noted above, $[z]_\Phi = [y]_\Phi$ and $[z]_\Phi = [x]_\Phi$ from which it follows that $[x]_\Phi = [y]_\Phi$. Therefore every pair of distinct equivalence classes has empty intersection and we deduce that the family of all equivalence classes is a partition, $\mathbf{P}(\Phi)$, of the set A.

There are some very interesting examples of equivalence relations. First let M be the set of all sequences $\mathbf{s} = (x_n)_{n\in\mathbb{N}}$ of rational numbers. Consider the relation Φ on M defined by the rule: $(\mathbf{s}, \mathbf{r}) \in \Phi$ if and only if

$$\lim_{n\to\infty} (x_n - y_n) = 0.$$

Here $\mathbf{r} = (y_n)_{n\in\mathbb{N}}$. It it easy to see that Φ is an equivalence relation.

For another example, let $M = [0, 1]$. Define a relation P on M by $(x, y) \in P$ if and only if $x - y$ is a rational number. It it easy to see that P is an equivalence relation.

There is one more important **example**.

Let m be a fixed natural number. Two integers are called congruent modulo m if $a - b$ is divisible by m, which we denote by $a \equiv b\,(\mathbf{mod}\ m)$. This congruence relation is easily shown to be an equivalence relation, which we shall consider in detail later.

We often denote equivalence relations using symbols such as $\cong$, $\equiv$, $\approx$, $\sim$, and others.

Exercise Set 1.5

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

1.5.1. On the set $G = \mathbb{Z} \times \{-1, 1\}$ we define an operation $*$ by the rule $(m, a) * (n, b) = (m + an, ab)$. Is this operation associative? Commutative? Is there an identity element? Which elements have inverses?

1.5.2. On a set of four elements define a commutative, associative binary operation having an identity element.

1.5.3. On the set $\mathbb{Z}$ define an operation $\perp$ by the rule $a \perp b = a^2 + b^2$, for $a, b \in \mathbb{Z}$. Is this operation associative? Commutative? Is there an identity element?

1.5.4. On the set $\mathbb{R}$ define an operation $\bullet$ by the rule $a \bullet b = a + b + ab$. Prove that

(i) $a \bullet (b \bullet c) = (a \bullet b) \bullet c$ for all $a, b, c \in \mathbb{R}$.

(ii) $a \bullet b = b \bullet a$ for all $a, b \in \mathbb{R}$.

(iii) if $a \neq -1$, then $a \bullet b = a \bullet c$ if and only if $b = c$.

Is there an identity element for this operation? Which elements have inverses?

1.5.5. On the set $\mathbb{R} \times \mathbb{R}$ define an operation $\bullet$ by the rule $(a,b) \bullet (c,d) = (ac-bd, bc+ad)$. Is this operation associative? Commutative? Is there an identity element?

1.5.6. Let $M = \{e,a,b,c\}$. Define a binary algebraic operation on M which is commutative, associative, and for which an identity element exists, but not every element has an inverse.

1.5.7. Let $M = \{e,a,b,c\}$. Define on M a binary algebraic operation which is commutative, associative, and for which there is an identity element, and every element has an inverse.

1.5.8. For $a,b \in \mathbb{R}$ define $a \simeq b$ to mean that $ab = 0$. Prove or disprove each of the following:

(a) The relation $\simeq$ is reflexive.
(b) The relation $\simeq$ is symmetric.
(c) The relation $\simeq$ is transitive.

1.5.9. For $a,b \in \mathbb{R}$ define $a \simeq b$ to mean that $ab \neq 0$. Prove or disprove each of the following:

(a) The relation $\simeq$ is reflexive.
(b) The relation $\simeq$ is symmetric.
(c) The relation $\simeq$ is transitive.

1.5.10. Fractions are numbers of the form $\frac{a}{b}$ where a and b are whole numbers and $b \neq 0$. Fraction equality is defined by $\frac{a}{b} = \frac{c}{d}$ if and only if $ad = bc$. Determine whether fraction equality is an equivalence relation.

1.5.11. Let $a,b \in \mathbb{N}$. Show that the operation $a * b = a^b + b^a$ is not associative on the set of natural numbers. Is it a commutative operation?

1.5.12. Let M be the set of sequences $\mathbf{s} = (x_n)_{n\in\mathbb{N}}$ of rational numbers and let also $\mathbf{r} = (y_n) \in M$. Prove that the relation Φ defined by $\mathbf{s}\,\Phi\,\mathbf{r}$ if and only if $\lim_{n\to\infty}(x_n - y_n) = 0$ is an equivalence relation on M. Write the first few terms of a sequence defining π.

1.5.13. Let m be a fixed natural number. If $a,b \in \mathbb{Z}$ then write $a \equiv b \pmod{m}$ if and only if m divides $b-a$. Prove that $\equiv$ is an equivalence relation on the set $\mathbb{Z}$. Write the equivalence class of $0 \pmod 7$ and the equivalence class of $0 \pmod 5$. Are these the same?

1.5.14. Prove that the relation "has the same image" on the elements of the domain of a function is an equivalence relation.

1.5.15. Define binary operations $\blacktriangledown, \blacktriangle$ and $\blacksquare$ on $\mathbb{Q}$ by the rules: $a\blacktriangledown b = a - b + ab, a\blacktriangle b = \frac{1}{2}(a+b+ab)$, $a\blacksquare b = \frac{1}{3}(a+b)$. Of these operations which are associative? Which are commutative? Which have an identity element?

1.5.16. Define a binary operation $\blacktriangledown$ on $\mathbb{R}$ by the rule: $a\blacktriangledown b = pa + qb + r$. For which fixed p, q, r, is this operation associative? For which values of p, q, r is the operation commutative? For which values of p, q, r is there an identity element?

1.5.17. Let $\mathbb{Q}^*$ be the set of all nonzero rational numbers. Which of the following properties hold for the operation of division:

(1) $a \div b = b \div a$;
(2) $(a \div b) \div c = a \div (b \div c)$;
(3) $((a \div b) \div c) \div d = a \div (b \div (c \div d))$;
(4) if $a \div b = a \div c$, then $b = c$;
(5) if $b \div a = c \div a$, then $b = c$.

1.5.18. For $a, b \in \mathbb{R}$ define $a \simeq b$ to mean that $|a - b| < 7$. Prove or disprove each of the following:

(a) The relation $\simeq$ is reflexive.
(b) The relation $\simeq$ is symmetric.
(c) The relation $\simeq$ is transitive.

1.5.19. For points $(a, b), (c, d) \in \mathbb{R}^2$ define $(a, b) \simeq (c, d)$ to mean that $a^2 + b^2 = c^2 + d^2$.

(a) Prove that $\simeq$ is an equivalence relation on $\mathbb{R}^2$.
(b) List all elements in the set $\{(x, y) \in \mathbb{R}^2 \mid (x, y) \simeq (0, 0)\}$.
(c) List five distinct elements in the set $\{(x, y) \in \mathbb{R}^2 \mid (x, y) \simeq (1, 0)\}$.

1.5.20. Two $n \times n$ matrices A and B are said to be similar if there exists an invertible $n \times n$ matrix P such that $P^{-1}AP = B$. Show that similarity is an equivalence relation on $\mathbf{M}_n(\mathbb{R})$.

2

NUMBERS

2.1 SOME PROPERTIES OF INTEGERS: MATHEMATICAL INDUCTION

In this chapter, we will consider some basic properties of the set $\mathbb{Z}$ of integers. This set contains the set of natural or counting numbers, $\mathbb{N} = \{1, 2, 3, 4, 5, \dots\}$, the set of whole numbers, $\mathbb{N}_0 = \{0, 1, 2, 3, 4, 5, \dots\}$, and the set of negative integers $\{\dots, -5, -4, -3, -2, -1\}$. Clearly $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

We begin by recalling the definition of the absolute value of an integer.

Let n be an integer. Then the absolute value of n is

$$|n| = \begin{cases} n, \text{if } n \in \mathbb{N}_0, \\ -n, \text{ if } n \notin \mathbb{N}_0. \end{cases}$$

You have significant experience working with integers, and in this section, we remind you of their main properties. First we discuss the operation of addition on the set $\mathbb{Z}$ which satisfies the following basic properties:

1. Addition *is commutative*, so for all integers $m, n \in \mathbb{Z}$,

$$m + n = n + m.$$

An Introduction to Essential Algebraic Structures, First Edition. Martyn R. Dixon, Leonid A. Kurdachenko and Igor Ya. Subbotin.

2. Addition *is associative*, so for all integers $k, m, n \in \mathbb{Z}$ we have

$$(k+m)+n = k+(m+n).$$

3. $\mathbb{Z}$ contains the number 0, a *zero element for addition*, which means that, for all $n \in \mathbb{Z}$, $0+n = n+0 = n$.
4. For each $n \in \mathbb{Z}$ there is an *additive inverse*, $-n \in \mathbb{Z}$, which satisfies

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

Next, we give the basic properties of the operation of multiplication on $\mathbb{Z}$. As usual if $m, n \in \mathbb{Z}$ then we write $m \cdot n$ as mn.

1. Multiplication *is commutative*, so for all integers $m, n \in \mathbb{Z}$,

$$mn = nm.$$

2. Multiplication *is associative*, so for all integers $k, m, n \in \mathbb{Z}$ we have

$$(km)n = k(mn).$$

3. $\mathbb{Z}$ contains the number 1, an *identity element for multiplication*, which means that, for all $n \in \mathbb{Z}$, $1n = n1 = n$.
4. *Zero property* of multiplication: For each $n \in \mathbb{Z}$,

$$n0 = 0n = 0.$$

The two operations of addition and multiplication are connected by the distributive law which asserts that, for all integers $k, m, n \in \mathbb{Z}$,

$$k(m+n) = km+kn.$$

This law is called the *left distributive law*. However, the properties of addition and multiplication listed above imply that the right distributive property also holds. Indeed, by the commutativity of multiplication, we can write $k(m+n) = (m+n)k$, $km = mk, kn = nk$, and the left distributive property directly implies *the right distributive property*

$$(m+n)k = mk+nk.$$

The set $\mathbb{Z}$ contains the subset $\mathbb{N}$ of all counting or natural numbers. We next note some properties of this set.

1. $\mathbb{N}$ is closed under addition, which means that the sum of every pair of natural numbers is a natural number.
2. $\mathbb{N}$ is closed under multiplication, which means that the product of every pair of natural numbers is a natural number.
3. The law of trichotomy holds, which means that, for each $n \in \mathbb{Z}$, one and only one of the following statements is true:
   (a) $n \in \mathbb{N}$,
   (b) $n = 0$, or
   (c) $-n \in \mathbb{N}$.

One very useful method of proof which is often employed when proving statements about the set of natural numbers is the Principle of Mathematical Induction. The main idea here is as follows. Let k be a natural number and suppose that $P(n)$ is a statement, or assertion, about the natural number n that is known to be true whenever $0 \leq n < k$. Suppose that it is then possible to prove that $P(k)$ is also valid. Then the Principle of Mathematical Induction asserts that $P(n)$ is valid for all $n \in \mathbb{N}_0$. In practice, it is necessary to check that the assertion $P(r)$ definitely holds for some small value of $r \in \mathbb{N}_0$. We then make the *induction hypothesis*, an assumption that $P(n)$ is true for all $r \leq n < m$, and use this to prove that $P(m)$ is true. The Principle of Mathematical Induction then asserts that $P(n)$ is true for all $n \geq r$.

Some care must be taken here. A key step in a proof by mathematical induction is to check that $P(n)$ holds for an appropriate small value of n (the so-called basis of the induction; often the basis is 0 or 1, but need not be so). If the verification of this step is neglected, we may "prove" a statement that is totally incorrect. For example, the statement "all people have the same eye color" could be "proved" in the following way. Let $P(n)$ denote the statement that in every set of n people, all n people have the same eye color. Indeed, for $n = 1$ the statement $P(1)$ is true. Suppose that $P(k)$ is true and consider a set S consisting of $k+1$ people. Then $S = M \cup \{x\}$, where M consists of k people. By the induction hypothesis, all people in the set M have the same eye color. Let $y \in M$ and $T = S \backslash \{y\}$. Then T has k people so all people in T have the same eye color. In particular, x and each person of T have the same eye color. This means that every person from M has the same eye color as x so that all people in S have the same eye color. Hence the assertion $P(k+1)$ is true.

The problem with this "proof" is that the first meaningful statement for the assertion that "all people have the same eye color" must deal with at least two people, the case $n = 2$. The statement $P(2)$, that all pairs of people have the same eye color, is of course false, so we might say that the induction process never starts. Usually, the problem itself will give you a hint concerning where the induction process should start.

We now illustrate the Principle of Mathematical Induction using some standard examples. First we prove the well-known equation

$$1+3+5+\cdots+(2n-1)=n^2 \text{ for all natural numbers } n. \tag{2.1}$$

It is clear that if $n=1$, then this equation is true since $1^2=1$ and $2\cdot 1-1=1$ also. To proceed to the inductive step, we suppose that we have already proved that Equation (2.1) is valid for all natural numbers $n \leq k$. In particular we have,

$$1+3+\cdots+(2k-1)=k^2. \tag{2.2}$$

We now use this statement to prove that Equation (2.1) holds in the case when $n=k+1$. We have, using Equation (2.2),

$$\begin{aligned} 1+3+\cdots+(2k-1)+(2k+1) &= [1+3+\cdots+(2k-1)]+(2k+1) \\ &= k^2+(2k+1)=(k+1)^2. \end{aligned}$$

This is Equation (2.1) with n replaced by $k+1$. Hence by the Principle of Mathematical Induction, we conclude that for any $n \in \mathbb{N}$ the equation $1+3+\cdots+(2n-1)=n^2$ is true for all natural numbers n.

For our next example, we prove that n^3+2n is divisible by 3, for all $n \in \mathbb{N}_0$.

We note that when $n=0$, $n^3+2n=0$, which is divisible by 3 so the induction starts. For the induction hypothesis, we assume that n^3+2n is divisible by 3, for all $n \leq k$, and proceed to prove that this statement is true for $n=k+1$. To this end consider $(k+1)^3+2(k+1)$. We have

$$\begin{aligned} (k+1)^3+2(k+1) &= (k^3+3k^2+3k+1)+(2k+2) \\ &= (k^3+2k)+(3k^2+3k+3) \\ &= (k^3+2k)+3(k^2+k+1). \end{aligned}$$

Here k^3+2k is divisible by 3, by the induction hypothesis, and $3(k^2+k+1)$ is also divisible by 3. Thus 3 divides $(k+1)^3+2(k+1)$ so our assertion holds for $n=k+1$ and the Principle of Mathematical Induction implies that the assertion is true for all $n \in \mathbb{N}$.

As a further illustration of the power of the method, we shall prove the well-known binomial theorem that for all real numbers x, y and for all natural numbers n,

$$\begin{aligned} (x+y)^n = x^n &+ C_1^n x^{n-1}y + C_2^n x^{n-2}y^2 \\ &+\cdots+C_k^n x^{n-k}y^k+\cdots+y^n. \end{aligned} \tag{2.3}$$

Here the integers C_k^n are the binomial coefficients that we compute by the formula

$$C_k^n = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\dots(n-k+1)}{1\cdot 2\dots(k-1)k},$$

where we interpret C_k^n to be 0 if $k > n$.

We will prove the formula (2.3) by induction on n.

Clearly, Equation (2.3) is valid for $n = 1, 2$ since $C_1^1 = C_0^1 = 1 = C_0^2 = C_2^2$ and $C_1^2 = 2$. For the induction hypothesis, we suppose that Equation (2.3) is true for all $n \le m$ and consider $(x+y)^{m+1}$. We have

$$\begin{aligned}
(x+y)^{m+1} &= (x+y)^m(x+y)\\
&= (x^m + C_1^m x^{m-1}y + C_2^m x^{m-2}y^2 + \cdots\\
&\quad + C_k^m x^{m-k}y^k + \cdots + y^m)(x+y)\\
&= x^m(x+y) + C_1^m x^{m-1}y(x+y) + \cdots\\
&\quad + C_k^m x^{m-k}y^k(x+y) + \cdots + y^m(x+y)\\
&= x^{m+1} + x^m y + \cdots + C_{k-1}^m x^{m+2-k}y^{k-1} + C_{k-1}^m x^{m+1-k}y^k\\
&\quad + C_k^m x^{m+1-k}y^k + C_k^m x^{m-k}y^{k+1} + \cdots + xy^m + y^{m+1}.
\end{aligned}$$

Combining like terms, we see that the term $x^{m+1-k}y^k$ has the coefficient

$$\begin{aligned}
C_{k-1}^m + C_k^m &= \frac{m!}{(k-1)!(m-k+1)!} + \frac{m!}{k!(m-k)!}\\
&= \frac{m!}{(k-1)!(m-k)!}\cdot\left(\frac{1}{m-k+1} + \frac{1}{k}\right)\\
&= \frac{m!}{(k-1)!(m-k)!}\,\frac{m+1}{k(m-k+1)} = \frac{(m+1)!}{k!(m+1-k)!} = C_k^{m+1}.
\end{aligned}$$

It follows that $(x+y)^{m+1}$ has the stated form and the result now follows by the Principle of Mathematical Induction.

Exercise Set 2.1

In each of the following questions, explain your reasoning by giving either a proof of your assertion or a counterexample.

2.1.1. Prove that, for each natural number n, $2 + 2^2 + 2^3 + 2^4 + \cdots + 2^n = 2^{n+1} - 2$.

2.1.2. Prove that $4n < 2^n$ for all natural numbers $n \ge 5$.

2.1.3. Show that $1+4+9+\cdots+n^2 = \dfrac{n(n+1)(2n+1)}{6}$.

2.1.4. Prove that $1+2^3+3^3+\cdots+n^3 = (\frac{n(n+1)}{2})^2$ for each positive integer n.

2.1.5. Prove that $2^n > 2n+1$ for each integer $n \geq 3$.

2.1.6. Prove that $(k!)^2 \geq k^k$ for all positive integers k.

2.1.7. Prove that $1 \cdot 1! + 2 \cdot 2! + 3 \cdot 3! + \cdots + n \cdot n! = (n+1)! - 1$ for each positive integer n.

2.1.8. Prove that n different lines lying in the same plane and intersecting in the same point divide this plane into $2n$ regions.

2.1.9. Show that

$$1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \frac{1}{\sqrt{4}} + \cdots + \frac{1}{\sqrt{n}} \leq 2\sqrt{n}$$

2.1.10. Show that $2!4!\ldots(2n)! \geq ((n+1)!)^n$

2.2 DIVISIBILITY

In this section, we discuss results connected with division properties of the integers and give the appropriate proofs.

Let us begin with an elementary **example** that of dividing 674 by 23. The first step in the long division process is to find an integer n such that $23n$ is less than or equal to 67, but $23(n+1)$ is larger than 67. It is easy to see that this integer is $n = 2$. So $23n = 46$. Now $67 - 46 = 21$, and we next consider the integer 214 ($674 - 460 = 214$). Again, we find an integer m such that $23m$ is less than or equal to 214, but $23(m+1)$ is larger than 214. This integer is $m = 9$, and $23m = 207$. The integer $214 - 207 = 7 < 23$ is the remainder, the quotient is 29, and we note that $674 = 29 \cdot 23 + 7$.

We usually write this process in the following familiar manner,

$$\begin{array}{r} 29 \\ 23\overline{)674} \\ \underline{46} \\ 214 \\ \underline{207} \\ 7 \end{array}$$

This means that for the two given integers $a = 674$ and $b = 23$, we found integers $q = 29$ and $r = 7$, such that $a = bq + r$. In other words, $674 = 29 \cdot 23 + 7$.

In the case when $b < 0$, let us say $b = -23$, we need to make some adjustments regarding the quotient. Indeed, $674 = (-29)(-23) + 7$. Note that although the quotient becomes negative, the remainder is still positive.

It is clear that the cases when a and b are negative, as, for example, $-674 = 29(-23) - 7$, or a is negative and b is positive ($-674 = (-29)23 - 7$), can be easily reduced to the two cases considered above. However, if we want to obtain a positive reminder, we need to add and subtract 23 on the right-hand side of the equation. For example, $-674 = 29(-23) - 23 + 23 - 7 = 30(-23) + 16$.

The following basic result is just a generalization of this process of long division. The reader can skip the proof. However, we make the general remark that understanding a proof often makes it easier to understand the statement of a theorem.

Theorem 2.2.1. *Let $a, b \in \mathbb{Z}$ and $b \neq 0$. Then there are integers q, r such that $a = bq + r$ and $0 \leq r < |b|$. The pair (q, r) having this property is unique.*

Proof. We split the proof into a series of cases. In the case when $|b| > |a|$, but $a > 0$ then $a = b \cdot 0 + a$ so take $q = 0$ and $r = a$. If $|b| > |a|$, but $a, b < 0$, then $b < a$ and since $b + |b| = 0$, we have $0 < a - b < |b|$ Thus $a = b \cdot 1 + (a - b)$ and we take $q = 1, r = a - b$. If $|b| > |a|$, but $a < 0 < b$, then $|b| - b = 0$ and $-a < b$. Then $a = b(-1) + (a + b)$ and we may take $q = -1, r = a + b$.

So we may assume that $|b| \leq |a|$. Suppose first that $b > 0$. There exists a positive integer n such that $nb > a$, so that $nb - a > 0$. Let

$$M = \{x \mid x \in \mathbb{N}_0 \text{ and } x = nb - a \text{ for some } n \in \mathbb{N}_0\}.$$

We have proved that M is not empty. If $0 \in M$, then $tb - a = 0$ for some positive integer t, and $a = bt$. In this case, we can put $q = t, r = 0$, so we may next suppose that $0 \notin M$. Since M is a subset of $\mathbb{N}_0$, M contains a least element $x_0 = bk - a$. Since $0 \notin M$, we have $x_0 \neq 0$. If $x_0 > b$, then $x_0 - b \in \mathbb{N}_0$, and $x_0 - b = b(k - 1) - a \in M$. We have $x_0 - (x_0 - b) = b > 0$, so $x_0 > (x_0 - b)$, and we obtain a contradiction to the choice of x_0. Thus $0 < bk - a \leq b$, which means that $-b < b(k - 1) - a \leq 0$. It follows that $0 \leq a - b(k - 1) < b$. Let $q = k - 1$, $r = a - bq$. Then $a = bq + r$ and $0 \leq r < b = |b|$.

Suppose next that $b < 0$. Then $-b > 0$, and by the previous argument there are integers m, r such that $a = (-b)m + r$ where $0 \leq r < -b = |b|$. We set $q = -m$. Then $a = bq + r$.

Finally, we prove the uniqueness of the pair q, r. Suppose also that $a = bq_1 + r_1$ where $0 \leq r_1 < |b|$. We have

$$bq + r = bq_1 + r_1 \text{ so } r - r_1 = bq_1 - bq = b(q_1 - q).$$

If $r = r_1$, then $b(q_1 - q) = 0$ and since $b \neq 0$, then $q_1 - q = 0$, so that $q_1 = q$. Therefore, assume that $r_1 \neq r$. Then either $r > r_1$ or $r < r_1$. If $r > r_1$, then

$$0 < r - r_1 < |b| \text{ so } 0 < |r - r_1| < |b|.$$

The equation $r - r_1 = b(q_1 - q)$ implies that $|r - r_1| = |b|\,|q - q_1|$. This shows that $|b| \leq |r - r_1|$, which is a contradiction.

If we suppose that $r < r_1$, then by the same argument, we again reach a contradiction. Hence $r = r_1$ and $q = q_1$. This completes the proof.

Definition 2.2.2. *Let $a, b \in \mathbb{Z}$ and $b \neq 0$. Then $a = bq + r$ where $0 \leq r < |b|$. The integer r is called a residue or a remainder. If $r = 0$, so that $a = bq$, then we say that b is a divisor of a, or that b divides a, or that a is divisible by b. We will write this symbolically as $b \mid a$.*

Of course if $|b| > |a|$, then $q = 0$ and $r = a$.

From this definition, we see that $a = a \cdot 1 = (-a)(-1)$, so ± 1 and $\pm a$ are divisors of a. If b is a divisor of a, then there exists an integer c such that $a = bc = (-b)(-c)$, whence $-b$ is also a divisor if a.

It is important to mention the intuitively clear fact that 1 and -1 are the only divisors of 1.

The following basic well-known properties of divisibility are useful and fairly intuitive. Their proofs are simple and clear, and we omit them here.

Theorem 2.2.3. *Let $a, b, c \in \mathbb{Z}$. Then the following properties of divisibility hold.*

(i) *If $a \mid b$ and $b \mid c$, then $a \mid c$.*
(ii) *If $a \mid b$, then $a \mid bc$.*
(iii) *If $a \mid b$, then $ac \mid bc$.*
(iv) *If $c \neq 0$ and $ac \mid bc$, then $a \mid b$.*
(v) *If $a \mid b$ and $c \mid d$, then $ac \mid bd$.*
(vi) *If $a \mid b$ and $a \mid c$, then $a \mid (bk + cm)$ for every $k, m \in \mathbb{Z}$.*

Definition 2.2.4. *Let $a, b \in \mathbb{Z}$. An integer d is called the greatest common divisor of a and b (and we write $d = \boldsymbol{GCD}(a, b)$), if it satisfies the conditions:*

(**GCD 1**) *d divides both a and b;*
(**GCD 2**) *if c divides a and c divides b, then c divides d.*

Since $0 = 0 \cdot a$, for each integer $a \neq 0$, it follows that $a \mid 0$ and Definition 2.2.4 implies that the greatest common divisor of 0 and a is a. Since 0 is divisible by

every integer, we must be careful with $\mathbf{GCD}(0,0)$ and, accordingly, we define $\mathbf{GCD}(0,0) = 0$.

Clearly if d is a greatest common divisor of a and b, then $-d$ is also a greatest common divisor of a and b. It is fairly easy to see that there are no other greatest common divisors for a and b.

The expression "d is the greatest common divisor of a and b" means that $|d|$ is greatest in absolute value of all the divisors of both numbers. For example, for 12 and 30, the numbers $-6,-3,-2,-1,1,2,3,6$ are common divisors, but 6 and -6 are the only greatest common divisors. Notice that -6 is the least common integer (by value) that is a divisor of 12 and 30.

It is important to investigate when a greatest common divisor exists. The next result shows that each pair of integers has such a **GCD**.

Theorem 2.2.5. *Let a,b be arbitrary integers. Then the greatest common divisor of a and b exists.*

Proof. Clearly $\mathbf{GCD}(a,0) = a$, so we may assume that $a,b \neq 0$. Let

$$M = \{ax+by \mid x,y \in \mathbb{Z}\}.$$

We observe that $a = a \cdot 1 + b \cdot 0 \in M$, so M is not empty. Furthermore, if $ax+by \in M$, but $ax+by \notin \mathbb{N}$, then

$$-(ax+by) = a(-x)+b(-y) \in M \cap \mathbb{N},$$

so that $M \cap \mathbb{N} \neq \emptyset$. Hence $M \cap \mathbb{N}$ has a least element, which we denote by d. Let $d = am+bn$, where $m,n \in \mathbb{Z}$. Suppose that d is not a divisor of a. By Theorem 2.2.1, $a = dq+r$ where $0 < r < d$. Then

$$r = a - dq = a - (am+bn)q = a(1-mq)+b(-nq) \in M \cap \mathbb{N},$$

and we obtain a contradiction to the choice of d. This contradiction shows that $r = 0$ and hence d divides a. Similarly, we can prove that d divides b.

Next let c be a common divisor of a and b. Then $a = ck$ and $b = cl$, where $k,l \in \mathbb{Z}$. We have

$$d = am+bn = (ck)m+(cl)n = c(km)+c(ln) = c(km+ln).$$

Since $m,l \in \mathbb{Z}$, c is a divisor of d, so that d satisfies the condition (**GCD 1**) and (**GCD 2**). This completes the proof.

From the proof of the previous theorem, we can extract the following consequence.

Corollary 2.2.6. *Let a,b be arbitrary integers and let $d = \mathbf{GCD}(a,b)$. Then there are integers m,n such that $d = am+bn$.*

We say that the integers a,b are *relatively prime*, or *coprime*, if $\mathbf{GCD}(a,b) = \pm 1$. The following result has many important applications in algebra and number theory.

Corollary 2.2.7. *Let a,b be integers. Then a and b are relatively prime if and only if there are integers m,n such that $1 = am+bn$.*

Proof. If a and b are relatively prime, then $\mathbf{GCD}(a,b) = 1$, and Corollary 2.2.6 proves the existence of m,n. Conversely, suppose that there are integers m,n such that $1 = am+bn$. Let $d = \mathbf{GCD}(a,b)$. Then $a = da_1, b = db_1$ for certain integers a_1, b_1. We have

$$1 = (da_1)m + (db_1)n = d(a_1m + b_1n).$$

Since 1 and -1 are the only integer divisors of 1, we deduce that $1 = |d|$, and hence, a and b are coprime.

If $d = \mathbf{GCD}(a,b)$, then, by Corollary 2.2.6, $d = am+bn$, for certain integers m,n. Since there are integers a_1, b_1 such that $a = da_1, b = db_1$ we have

$$d = da_1m + db_1n \text{ so } 1 = a_1m + b_1n.$$

This proves the following result, which is intuitively obvious: if we divide out the greatest common divisor of two integers, then the corresponding quotients have nothing left in common.

Corollary 2.2.8. *Let a,b be integers, $d = \mathbf{GCD}(a,b)$ and $a = da_1, b = db_1$. Then a_1 and b_1 are relatively prime.*

Corollary 2.2.9. *Let a,b,c be integers.*

(i) *If a divides bc and a,b are relatively prime, then a divides c;*
(ii) *If a,b are relatively prime and a,c are also relatively prime, then a and bc are relatively prime;*
(iii) *If a,b divide c and a,b are relatively prime, then ab divides c.*

Proof.

(i) By Corollary 2.2.7, there are integers m,n such that $am+bn = 1$. Then

$$\begin{aligned} c &= c(am+bn) = c(am) + c(bn) \\ &= (ca)m + (cb)n = (ac)m + (bc)n = a(cm) + (bc)n. \end{aligned}$$

However, $a|bc$ also, so Theorem 2.2.3(vi) shows that a divides $a(cm)+(bc)n$ and hence a divides c.

(ii) By Corollary 2.2.7, there are integers m,n,k,t such that $am+bn=1$ and $ak+ct=1$. Then

$$1=(am+bn)(ak+ct)=a(mak+mct+bnk)+(bc)(nt),$$

and, again using Corollary 2.2.7, we deduce that a and bc are relatively prime.

(iii) By Corollary 2.2.7, there are integers m,n such that $am+bn=1$. We have $c=au$ and $c=bv$ for certain integers u,v. Then

$$\begin{aligned} c&=c(am+bn)=c(am)+c(bn)=(bv)(am)+(au)(bn)\\ &=(ab)(vm)+(ab)(un)=ab(vm+un). \end{aligned}$$

Thus, $ab \mid c$.

Theorem 2.2.5 established the existence of a greatest common divisor for each pair of integers. However, it does not really provide a method for finding this greatest common divisor. The well-known Euclidean Algorithm is a commonly used procedure for doing this.

Let a,b be integers. If $a=0$, then $\mathbf{GCD}(a,b)=b$. Therefore, we can assume that $a,b\neq 0$. By Theorem 2.2.1, $a=bq_1+r_1$, where $0\leq r_1<|b|$. If $r_1\neq 0$, we have, again by Theorem 2.2.1, $b=r_1q_2+r_2$, where $0\leq r_2<r_1$. If $r_2\neq 0$, then $r_1=r_2q_3+r_3$, where $0\leq r_3<r_2$. If $r_3\neq 0$, then we can continue this process. In general, if $r_j\neq 0$, then we continue the division process by dividing r_j by r_{j+1} and note that at each step, $0\leq r_{j+1}<r_j$. Therefore, this process must terminate in finitely many steps, which means that at some point r_k must be 0. Thus we obtain the following chain.

$$\begin{aligned} a&=bq_1+r_1,\\ b&=r_1q_2+r_2,\\ r_1&=r_2q_3+r_3,\\ &\dots\\ r_{k-3}&=r_{k-2}q_{k-1}+r_{k-1},\\ r_{k-2}&=r_{k-1}q_k+r_k,\\ r_{k-1}&=r_kq_{k+1}+0. \end{aligned} \tag{2.4}$$

We have

$$r_{k-2}=r_{k-1}q_k+r_k=r_kq_{k+1}q_k+r_k=r_k(q_{k+1}q_k+1),$$

so that r_k divides r_{k-2}.

Furthermore,

$$r_{k-3} = r_{k-2}q_{k-1} + r_{k-1} = r_k(q_{k+1}q_k + 1)q_{k-1} + r_k q_{k+1}$$
$$= r_k(q_{k+1}q_k q_{k-1} + q_{k-1} + q_{k+1}),$$

so that r_k divides r_{k-3}. By continuing up the chain (2.4), we deduce, informally, that r_k divides a and b. A more formal proof would use mathematical induction.

Now let u be an arbitrary common divisor of a and b. The equation $r_1 = a - bq_1$ shows that u is a divisor of r_1. The next equation, $r_2 = b - r_1 q_2$, shows that u is a divisor of r_2. By coming down the chain (2.4), we deduce that u is a divisor of r_k. This means that r_k is a greatest common divisor of a and b.

Corollary 2.2.6 shows that there are integers x, y such that $r_k = ax + by$. The Euclidean Algorithm shows us how to find these integers x, y. Indeed we have

$$r_k = r_{k-2} - r_{k-1}q_k.$$

Also,

$$r_{k-1} = r_{k-3} - r_{k-2}q_{k-1},$$

so that

$$r_k = r_{k-2} - (r_{k-3} - r_{k-2}q_{k-1})q_k = r_{k-2} - r_{k-3}q_k + r_{k-2}q_{k-1}q_k$$
$$= r_{k-2}(1 + q_{k-1}q_k) - r_{k-3}q_k = r_{k-2}y_1 - r_{k-3}x_1, \text{ say.}$$

Using the further equation $r_{k-2} = r_{k-4} - r_{k-3}q_{k-2}$, we prove that $r_k = r_{k-3}y_2 - r_{k-4}x_2$. Continuing in this way and moving back along the chain in Equation (2.4), we finally obtain the equation $r_k = ax + by$. The values of x and y will then be evident.

Of course many standard computer programs will compute the greatest common divisor of two integers in an instant.

Example.

Let $a = 834$, $b = 154$.

Then $834 = 154 \times 5 + 64$, so $\mathbf{GCD}(834, 154) = \mathbf{GCD}(154, 64)$.

Next $154 = 64 \times 2 + 26$, and $\mathbf{GCD}(154, 64) = \mathbf{GCD}(64, 26)$.

Also $64 = 26 \times 2 + 12$, so $\mathbf{GCD}(64, 26) = \mathbf{GCD}(26, 12)$.

Then $26 = 12 \times 2 + 2$. Therefore, $\mathbf{GCD}(26, 12) = \mathbf{GCD}(12, 2) = 2$.

We have therefore shown that $\mathbf{GCD}(834, 154) = 2$.

We can also find the greatest common divisor of any two integers using their prime factorizations. Thus $834 = 2 \times 3 \times 139, 154 = 2 \times 7 \times 11$, so $\mathbf{GCD}(834, 154) = 2$.

Exercise Set 2.2

In each of the following questions, explain your reasoning by giving either a proof of your assertion or a counterexample.

2.2.1. Prove that n^2+n is even for each positive integer n.

2.2.2. Prove that 3 divides n^3-n for each positive integer n.

2.2.3. Prove that the product of three consecutive positive integers is divisible by 6.

2.2.4. Prove that 8 divides n^2-1 for each odd positive integer n.

2.2.5. Prove that the product of 5 consecutive positive integers is divisible by 120.

2.2.6. Prove that 133 divides $11^{n+2}+12^{2n+1}$ for each integer $n \geq 0$.

2.2.7. Find a two digit number n satisfying the following criteria: When n is divided by the sum of its digits the quotient is 4 and the remainder is 3; when n is divided by the product of its digits the quotient is 3 and the remainder is 5

2.2.8. Prove that 24 divides $n^4+6n^3+11n^2+6n$ for each positive integer n.

2.2.9. Prove that 30 divides $k^{73}-k^{37}$ for all positive integers k.

2.2.10. Find all positive integers x such that 9 divides x^2+2x-3.

2.2.11. Prove Theorem 2.2.3.

2.2.12. Prove that if $a,b \in \mathbb{Z}$ and if $d = \mathbf{GCD}(a,b)$ then the only other greatest common divisor of a and b is $-d$.

2.2.13. Let $a,b \in \mathbb{Z}$ and let $d = \mathbf{GCD}(a,b)$. Prove that if $c \in \mathbb{Z}$ and there exist $u,v \in \mathbb{Z}$ such that $c = au+bv$, then d divides c.

2.2.14. Find the greatest common divisor d of a,b and find integers x,y such that $ax+by=d$ for the following:
(a) $a = 308, b = 455$, (b) $a = 780, b = 462$.

2.2.15. If $a,b \in \mathbb{Z}$, then the least common multiple of a and b, denoted by $\mathbf{LCM}(a,b)$ is the positive integer d such that $a|d, b|d$ and whenever $a|x, b|x$ for some $x \in \mathbb{Z}$ then $d|x$. Prove that $\mathbf{LCM}(a,b) = ab/\mathbf{GCD}(a,b)$ if $a,b > 0$.

2.2.16. Let n,k be positive integers. Let $n = kq+r$ where $0 \leq r < k$. Prove that $\mathbf{GCD}(n,k) = \mathbf{GCD}(k,k-r)$.

2.2.17. Let a, b, c be nonzero integers. Prove that $\mathbf{GCD}(\mathbf{GCD}(a,b),c) = \mathbf{GCD}(a,\mathbf{GCD}(b,c))$.

2.2.18. Let n be an arbitrary natural number. Prove that $\mathbf{GCD}(11n+3, 4n+1) = 1$.

2.2.19. Let $a, b, c \in \mathbb{Z}$. Show that if $\mathbf{GCD}(a,c) = \mathbf{GCD}(b,c) = 1$ then $\mathbf{GCD}(ab,c) = 1$.

2.2.20. Let $x, y, z \in \mathbb{Z}$. Show that if x divides $y+z$ and if $\mathbf{GCD}(y,z) = 1$, then $\mathbf{GCD}(x,y) = \mathbf{GCD}(x,z) = 1$.

2.3 PRIME FACTORIZATION: THE FUNDAMENTAL THEOREM OF ARITHMETIC

In this section, we consider only nonnegative integers, the so-called whole numbers, and as usual we denote this set by $\mathbb{N}_0$.

Definition 2.3.1. *Let $a \in \mathbb{N}_0$. A divisor of a, different from 1 or a, that lies in $\mathbb{N}_0$ is called a proper divisor of a. The divisors 1 and a are called improper divisors of a. A nonzero natural number p is called prime if $p > 1$ and p has no proper positive divisors. An integer that is not prime is called composite.*

Example. The first few prime numbers are $2, 3, 5, 7, 11, 13, 17, \ldots$, and below we prove that the set of all primes is infinite. Clearly, every even number that is at least 4 is composite. It is an important problem to determine which integers are prime. One method for doing this was proposed by Eratosthenes (276–194 BC), an ancient Greek mathematician, using a technique now called The Sieve of Eratosthenes. This is a simple algorithm, which, in principle, will find all primes up to a specified whole number n and consists of the following steps:

1. Write a list of all natural numbers from 2 to the given number n.
2. Delete from this list all multiples of two (4, 6, 8, etc.), since these are not prime.
3. Delete from this list all remaining multiples of three (9, 15, etc.), since these are not prime.
4. Find in the list the next remaining prime number (5) and delete all numbers that are multiples of 5, and so on.

Note that the primes $2, 3, \ldots$ are not deleted in this process—in fact the primes are the only numbers remaining at the end of the process. The process

is continued until an integer larger than $\sqrt{n}$ is reached at which time all remaining numbers that have not been deleted will be prime. For if n is a natural number that is not prime, then it must have a prime factor less than $\sqrt{n}$, otherwise n would be a product of at least two natural numbers both strictly larger than $\sqrt{n}$, which is impossible. As a consequence, we need only delete multiples of primes that are less than or equal to $\sqrt{n}$.

However, in general there is no known efficient algorithm that allows us to generate new primes from ones already known in a finite number of steps. The discovery of a new prime is now greeted with some fanfare. Currently, the greatest known prime is $2^{57885161} - 1$, consisting of 17425170 digits which was discovered in August 2008.

Our next theorem is usually dubbed "The Fundamental Theorem of Arithmetic."

Theorem 2.3.2. *Let a be a natural number such that $a > 1$. Then a is a product of primes and this decomposition is unique up to the order of the factors. Furthermore,*

$$a = p_1^{k_1} ... p_m^{k_m}$$

where $k_1, ..., k_m$ are positive whole numbers, $p_1, ..., p_m$ are primes and $p_j \neq p_s$ whenever $j \neq s$.

Proof. We proceed by induction on a. If a is a prime (in particular, $a = 2, 3$), then the result follows. Suppose that a is a not prime and that every nonzero whole number u such that $1 < u < a$ decomposes as a product of primes. Then a has a proper divisor b, so $a = bc$ for some whole number c and clearly $b < a$. For the same reason, $c < a$. Then by the induction hypothesis, b and c can be written as products of primes, so that $a = bc$ has a similar decomposition. This proves the first part of the theorem.

Next let $a = p_1 \dots p_n$ where $p_1, \dots, p_n$ are primes. We shall prove the uniqueness of this decomposition by induction on n. Suppose that $a = q_1 ... q_t$ where $q_1, ..., q_t$ are primes. If $n = 1$, then $a = p_1$ is a prime. We have $p_1 = q_1 ... q_t$. Since q_1 is prime, $q_1 \neq \pm 1$. It follows that $q_2 ... q_t = \pm 1$, which means that $t = 1$ and $p_1 = q_1$. Suppose now that $n > 1$ and that our assertion is already proved for numbers that are products of fewer than n prime factors. We have $p_1 \dots p_n = q_1 \dots q_t$. Let $d =$**GCD**(p_1, q_1). Since p_1, q_1 are prime, either $d = 1$ or $q_1 = d = p_1$. In the latter case, it follows that $p_2 \dots p_n = q_2 \dots q_t$. By the induction hypothesis, we deduce that $n - 1 = t - 1$ and after some renumbering $q_j = p_j$, for $2 \leq j \leq n$. Then $n = t$ and $q_j = p_j$ for $1 \leq j \leq n$. If $d = 1$ then p_1 divides $q_2 \dots q_m$, by Corollary 2.2.9. By the same argument, either **GCD**$(p_1, q_2) = 1$ or $p_1 = q_2$. By repeating this type of argument enough times, we deduce that

there is a natural number k such that $p_1 = q_k$. By renumbering if necessary, we may suppose that $p_1 = q_1$. Then it follows that $p_2 \dots p_n = q_2 \dots q_t$, so that, as above $n = t$ and $q_j = p_j$, for $1 \leq j \leq n$.

This proves the uniqueness of the decomposition. Collecting together all equal prime factors, we arrive at the decomposition $a = p_1^{k_1} \dots p_m^{k_m}$, where $p_j \neq p_s$ whenever $j \neq s$.

The existence and uniqueness of the decomposition of integers into a product of prime factors was assumed as an obvious fact up to the end of the eighteenth century, so until that time mathematicians did not feel the need of such a result. The situation changed in 1801 when the great German mathematician Carl Frederick Gauss (1777–1855), known as the "Prince of Mathematicians," because of his many significant contributions to numerous fields, including number theory, clearly formulated and proved the Fundamental Theorem of Arithmetic.

The following absolutely brilliant proof that the set of all primes is infinite was published in Euclid's *Elements* (about 300 BC!).

Theorem 2.3.3. *The set of all primes is infinite.*

Proof. We suppose the contrary that the set of all primes $\{p_1, \dots, p_n\}$ is finite. Let

$$a = 1 + p_1 \dots p_n.$$

By Theorem 2.3.2, a is a product of primes so there is some prime p_i dividing a. However, p_i also divides $p_1 \dots p_n$ so p_i divides $a - p_1 \dots p_n$. But $a - p_1 \dots p_n = 1$ so p_i divides 1. Since $p_i > 1$ this is a contradiction, which proves the result.

It is important to observe that even though the set of primes is infinite, at the current moment there is one largest *known* prime.

Theorem 2.3.2 states that every natural number can be decomposed as a product of prime powers, but it does not show how this can be done and there is no good fast algorithm for doing this. This fact is used in cryptography, which is why the discovery of any new prime is of interest outside the scientific community.

We conclude this section with a short discussion concerning special types of prime. A prime number M_p is called a *Mersenne prime* if it is of the form $M_p = 2^p - 1$, where p is another prime. Such numbers were discussed in Euclid's *Elements*, but were named after Meren Mersenne (1588–1648) who made an intense study of them. By 1750, only seven such prime numbers had been discovered, corresponding to the primes $= 2, 3, 5, 7, 13, 17, 19$.

In 1751, Leonard Euler found the next Mersenne prime M_{31}. Before the computer era, only four additional Mersenne primes were found, namely, the prime numbers corresponding to the primes $p = 61, 89, 107, 127$. Since these numbers are huge e.g., $2^{11213} - 1$ is a Mersenne prime), computers have not significantly increased the list of such numbers.

The Fermat primes, denoted by F_n, have also been intensively studied. These are primes of the form $F_n = 2^{k(n)} + 1$, where $k(n) = 2^n$, and were introduced by the famous French mathematician Pierre de Fermat who listed the first five of them $F_0 = 3, F_1 = 5, F_2 = 17, F_3 = 257, F_4 = 65537$, all of which are prime. Fermat conjectured that all such numbers are prime. However, Euler showed that F_5 is not prime and, interestingly, no other Fermat primes have been discovered, despite intensive computer searches.

Exercise Set 2.3

In each of the following questions, explain your reasoning by giving either a proof of your assertion or a counterexample.

2.3.1. Find all positive integers x such that $x^2 + 2x - 3$ is a prime.

2.3.2. Prove that for all natural numbers x greater than 1 the number $x^2 + 5x - 6$ is composite.

2.3.3. Find all positive integers x such that $x^2 - 5x + 6$ is a prime.

2.3.4. Prove that the sum of five consecutive natural numbers is never prime.

2.3.5. What is the smallest natural number divisible by $1, 2, 3, 4, \ldots, 24, 25$?

2.3.6. Prove that if p is a prime greater than 5, then the number $111\ldots1$, consisting of $p - 1$ ones, is divisible by p.

2.3.7. Prime triples are three consecutive primes whose differences are 2. Find all of such triples.

2.3.8. Suppose that $2^n + 1$ is a prime where n is a positive integer. Prove that $n = 2^k$ for some positive integer k.

2.3.9. The well-known Goldbach conjecture asserts that each even integer greater than 2 can be expressed as the sum of two primes. Mathematicians have not been able to prove this conjecture yet, so it may even be false. Assuming that the conjecture is true, show that every odd whole number greater than 6 is the sum of three primes.

2.3.10. Prove that the sum of the first n consecutive natural numbers is not prime, when $n > 2$.

2.3.11. Let p be a prime. Let $\binom{p}{i} = p!/i!(p-i)!$ denote the binomial coefficient, for each i such that $0 < i < p$. Prove that $\binom{p}{i}$ is divisible by p. Use this to prove, by induction on $n \in \mathbb{N}$, Fermat's Little Theorem, namely that p divides $n^p - n$.

2.3.12. List all the positive primes less than 200.

2.3.13. Give three different examples of primes of the form $4k+1$, where k is a natural number. Prove that a product of two numbers of the form $4k+1$, where k is a natural number, is again of the form $4k+1$. Use the Principle of Mathematical Induction to generalize this to products of more than two primes.

2.3.14. Give three different examples of primes of the form $4k+3$, where k is a natural number. Prove that there are infinitely many primes of the form $4k+3$. [Hint: Suppose there are only finitely many primes $p_1, p_2, \ldots, p_r$ of this form and then show that $N = 4p_1p_2 \ldots p_r - 1$ is divisible by at least one prime of the form $4k+3$ and use this to obtain a contradiction.]

2.3.15. Prove that there are infinitely many primes of the form $6k+5$.

2.3.16. Let n be a natural number and let $\phi(n)$ denote the number of integers m such that $1 \leq m < n$ and such that m is relatively prime to n. Show that if p is a prime and $k \in \mathbb{N}$, then $\phi(p^k) = p^{k-1}(p-1)$. The function ϕ is called Euler's ϕ function or Euler's totient function.

2.3.17. Compute $\phi(236)$ and $\phi(935)$, given that if m, n are relatively prime then $\phi(mn) = \phi(m)\phi(n)$.

2.4 RATIONAL NUMBERS, IRRATIONAL NUMBERS, AND REAL NUMBERS

In this section, we consider the set $\mathbb{Q}$ of rational numbers. This set is an extension of the set of integers in which addition and multiplication possess the same properties, and division by all nonzero elements is also defined. Of course, we keep this extension minimal with respect to the properties we would like to keep. Most university courses pay no attention to the axiomatic construction of the sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$. One reason for this is the belief that the naive experience gained by students in high school is sufficient. Perhaps this is mostly true regarding the sets $\mathbb{N}$ and $\mathbb{Z}$, but the set $\mathbb{Q}$ has significantly different properties and characteristics. One important feature of the set $\mathbb{Q}$ is

the fact that a given rational number can be written in different ways. Rational numbers are actually equivalence classes. Treating a rational number as a number of parts of some unit makes it easy to understand a simple equation such as $\frac{3}{5} = \frac{6}{10} = \frac{9}{15}$.

To rigorously construct the set of real numbers requires some effort and is usually done in a Real Analysis course. To construct the rational numbers, much less effort is required. On the other hand, the construction of the set $\mathbb{Q}$ is a very natural place to illustrate important concepts like equivalence relations and partitions. Therefore, we here give a strict construction of the set of rational numbers.

To formally construct the set of rational numbers, we consider the Cartesian product $\mathbb{Z} \times (\mathbb{Z}\backslash\{0\})$ and a certain partition of this set. The elements of this partition are called *the rational numbers*.

This partition is natural once we recall that the way a fraction $\frac{a}{b}$ is defined is not unique. However, the fractions $\frac{a}{b}$ and $\frac{c}{d}$ are equal if and only if $ad = bc$. For **example**, $\frac{1}{3} = \frac{2}{6} = \frac{3}{9} = \cdots, 5 = \frac{5}{1} = \frac{10}{2} = \frac{15}{3} = \cdots$, so any rational number is a representative of a class of infinitely many fractions. In other words, with each rational number (fraction), we associate an infinite set of pairs of integers.

Definition 2.4.1. *The pairs (a,b) and (c,d), where a,b,c,d are integers and $b,d \neq 0$, are equivalent, if $ad = bc$. Let*

$$\frac{a}{b} = \{(c,d) \in \mathbb{Z} \times (\mathbb{Z}\backslash\{0\}) \mid ad = bc\}.$$

The set $\frac{a}{b}$ is called a fraction; a is called the numerator of the fraction and b is called the denominator of the fraction.

Certainly this relation is a reflexive one since $ab = ba$ for all $a,b \in \mathbb{Z}$. From the definition, it follows that if (a,b) and (c,d) are equivalent, then (c,d) and (a,b) are equivalent, so this relation of equivalence has the symmetric property. We note that $\frac{a}{b}$ is determined equally well by each of its elements; thus if $(k,n) \in \frac{a}{b}$, then $\frac{a}{b} = \frac{k}{n}$. In fact, let $(u,v) \in \frac{a}{b}$, so that $av = bu$. We have also $an = bk$. Then

$$(av)n = (bu)n \text{ and } (an)v = (bk)v.$$

It follows, using associativity and commutativity, that $b(un) = (bu)n = (av)n = b(kv)$. Since $b \neq 0$, $un = kv$, so the pairs (k,n) and (u,v) are equivalent and $(u,v) \in \frac{k}{n}$. This means that $\frac{a}{b} \subseteq \frac{k}{n}$. By the same argument, we also have $\frac{k}{n} \subseteq \frac{a}{b}$, so $\frac{a}{b} = \frac{k}{n}$.

Moreover, we may also deduce that if the pairs (u,v) and (a,b) are equivalent and the pairs (a,b) and (k,n) are also equivalent, then the pairs (u,v)

and (k,n) are equivalent, so this relation has the transitive property. Thus the relation given in Definition 2.4.1 is an equivalence relation on the set $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$.

Suppose that $(a/b) \cap (t/m) \neq \emptyset$ and let $(k,n) \in \frac{a}{b} \cap \frac{t}{m}$. As we saw earlier, $\frac{a}{b} = \frac{k}{n}$ and $\frac{t}{m} = \frac{k}{n}$, so $\frac{a}{b} = \frac{t}{m}$. On the other hand, $(a,b) \in \frac{a}{b}$ for each pair (a,b). Hence the set of all subsets $\frac{a}{b}$, where $(a,b) \in \mathbb{Z} \times (\mathbb{Z}\setminus\{0\})$ is a partition of $\mathbb{Z} \times (\mathbb{Z}\setminus\{0\})$.

Definition 2.4.2. *The set of rational numbers is defined to be the set*

$$\mathbb{Q} = \left\{ \frac{a}{b} \mid (a,b) \in \mathbb{Z} \times (\mathbb{Z}\setminus\{0\}) \right\}.$$

Definition 2.4.3. *The operations of addition and multiplication are defined on $\mathbb{Q}$ by the rules:*

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd};$$
$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

We must be sure that these operations are well defined, which in this case means that they are independent of the choice of element from the set $\frac{a}{b}$.

To do this, let $(k,n) \in \frac{a}{b}$, $(t,m) \in \frac{c}{d}$. This means that $an = bk$ and $cm = dt$. Then

$$\begin{aligned}(ad+bc)(nm) &= ad \cdot nm + bc \cdot nm = (an)(dm) + (cm)(bn) \\ &= bk \cdot dm + dt \cdot bn = (km+nt)bd.\end{aligned}$$

It follows that the pairs $(ad+bc, bd)$ and $(km+nt, nm)$ are equivalent, so that the addition is well defined.

Furthermore,

$$(ac)(nm) = (an)(cm) = bk \cdot dt = (bd)(kt),$$

which implies that the pairs (ac, bd) and (kt, nm) are equivalent. Therefore, multiplication is well defined also.

The basic properties of these operations are obtained next.

Theorem 2.4.4. *Let $a,b,c,d,u,v \in \mathbb{Z}$, where b,d,v are nonzero. The following properties hold.*

(i) $\frac{a}{b} + \frac{c}{d} = \frac{c}{d} + \frac{a}{b}$, *so addition is commutative;*
(ii) $(\frac{a}{b} + \frac{c}{d}) + \frac{u}{v} = \frac{a}{b} + (\frac{c}{d} + \frac{u}{v})$, *so addition is associative;*

(iii) $\frac{a}{b}+\frac{0}{d}=\frac{a}{b}$, *so the fraction* $\frac{0}{d}$ *is the zero element for addition;*

(iv) *The fraction* $\frac{-a}{b}$ *is an additive inverse to* $\frac{a}{b}$, *so every element of* $\mathbb{Q}$ *has an additive inverse;*

(v) $\frac{a}{b}\cdot\frac{c}{d}=\frac{c}{d}\cdot\frac{a}{b}$, *so multiplication is commutative;*

(vi) $\frac{a}{b}\cdot(\frac{c}{d}\cdot\frac{u}{v})=(\frac{a}{b}\cdot\frac{c}{d})\cdot\frac{u}{v}$, *so multiplication is associative;*

(vii) *The fraction* $\frac{d}{d}$ *is the multiplicative identity for each nonzero* $d\in\mathbb{Z}$;

(viii) *If* $a\neq 0$, *then* $\frac{a}{b}\cdot\frac{b}{a}$ *is the multiplicative identity, so every nonzero fraction has a reciprocal or multiplicative inverse;*

(ix) $\frac{a}{b}(\frac{c}{d}+\frac{u}{v})=\frac{a}{b}\cdot\frac{c}{d}+\frac{a}{b}\cdot\frac{u}{v}$, *so multiplication is distributive over addition.*

Proof.

(i) We have

$$\frac{a}{b}+\frac{c}{d}=\frac{ad+bc}{bd}=\frac{cb+da}{db}=\frac{c}{d}+\frac{a}{b}.$$

(ii) Next

$$\frac{a}{b}+\left(\frac{c}{d}+\frac{u}{v}\right)=\frac{a}{b}+\frac{cv+du}{dv}=\frac{a(dv)+(cv+du)b}{b(dv)}.$$

On the other hand,

$$\left(\frac{a}{b}+\frac{c}{d}\right)+\frac{u}{v}=\frac{ad+bc}{bd}+\frac{u}{v}=\frac{(ad+bc)v+(bd)u}{(bd)v}.$$

Since

$$\begin{aligned}a(dv)+(cv+du)b&=a(dv)+(cv)b+(du)b=a(dv)+b(cv)+b(du)\\&=(ad)v+(bc)v+(bd)u=(ad+bc)v+(bd)u,\end{aligned}$$

using commutativity, associativity, and distributivity in $\mathbb{Z}$ and, since $(bd)v=b(dv)$, (ii) follows.

(iii) Next

$$\frac{a}{b}+\frac{0}{d}=\frac{ad+b0}{bd}=\frac{ad}{bd}.$$

Since $a(bd)=(ab)d=(ba)d=b(ad)$, the fractions $\frac{a}{b}$ and $\frac{ad}{bd}$ coincide so $\frac{a}{b}+\frac{0}{d}=\frac{a}{b}=\frac{0}{d}+\frac{a}{b}$. We note that

$$\frac{0}{d}=\frac{0}{v}$$

for all nonzero $d,v\in\mathbb{Z}$, so $\frac{0}{d}$ is the zero element.

(iv) Also,

$$\frac{a}{b}+\frac{-a}{b}=\frac{ab+b(-a)}{b^2}=\frac{ab+(-ba)}{b^2}=\frac{ab-ba}{b^2}=\frac{0}{b^2}=\frac{-a}{b}+\frac{a}{b},$$

so $\frac{-a}{b}$ is the negative of $\frac{a}{b}$.

(v) For the commutative property of multiplication, we see

$$\frac{a}{b}\cdot\frac{c}{d}=\frac{ac}{bd}=\frac{ca}{db}=\frac{c}{d}\cdot\frac{a}{b}.$$

(vi) can be proved in a similar fashion to (v), using the associative property of multiplication in the set of integers.

(vii) To show that $\frac{d}{d}$ is the multiplicative identity, we note that

$$\frac{a}{b}\cdot\frac{d}{d}=\frac{ad}{bd}=\frac{d}{d}\cdot\frac{a}{b}$$

and since $(ad)b=(bd)a$, we have

$$\frac{ad}{bd}=\frac{a}{b}.$$

(viii) To see that $\frac{a}{b}$ has a multiplicative inverse we note that when $a\neq 0$, $\frac{b}{a}\in\mathbb{Q}$ and

$$\frac{a}{b}\cdot\frac{b}{a}=\frac{ab}{ba}=\frac{ab}{ab}=\frac{b}{a}\cdot\frac{a}{b}.$$

By (vii), $\frac{ab}{ab}$ is the identity element, so b/a is the multiplicative inverse of a/b.

(ix) For the distributive property, notice that

$$\frac{a}{b}\left(\frac{c}{d}+\frac{u}{v}\right)=\frac{a}{b}\cdot\left(\frac{cv+du}{dv}\right)=\frac{a(cv+du)}{bdv}=\frac{acv+adu}{bdv}.$$

On the other hand,

$$\frac{a}{b}\cdot\frac{c}{d}+\frac{a}{b}\cdot\frac{u}{v}=\frac{ac}{bd}+\frac{au}{bv}=\frac{ac(bv)+bd(au)}{bdbv}=\frac{b(acv+dau)}{b(dbv)}.$$

By (vii) we see,

$$\frac{acv+dau}{dbv} = \frac{b(acv+adu)}{b(dbv)},$$

so the distributive property follows. The proof is complete.

The existence of additive inverses allows us to define the operation of *subtraction* of fractions. We define the difference of two fractions $\frac{n}{k}$ and $\frac{c}{d}$ by

$$\frac{n}{k} - \frac{c}{d} = \frac{n}{k} + \left(-\frac{c}{d}\right) = \frac{nd+k(-c)}{kd} = \frac{nd-kc}{kd}.$$

It is easy to check that this operation satisfies the usual properties of subtraction, already obtained for integers.

Next we show how to embed $\mathbb{Z}$ into $\mathbb{Q}$. To this end, let $n,u,v \in \mathbb{Z}$ and let $u \neq 0$. Since $(nu)v = u(nv)$ the pairs (nu,u) and (nv,v) are equivalent, so that $\frac{nu}{u} = \frac{nv}{v}$. We define the mapping $\iota : \mathbb{Z} \longrightarrow \mathbb{Q}$, by $\iota(n) = \frac{nu}{u}$, where $n,u \in \mathbb{Z}$ and $u \neq 0$. The previous argument shows that ι is independent of the choice of u, so ι is well defined. Suppose that $n \neq k$, but $\frac{nu}{u} = \frac{kv}{v}$ where $0 \neq u,v \in \mathbb{Z}$. Then

$$n(uv) = (nu)v = u(kv) = (kv)u = k(vu) = k(uv).$$

Since $uv \neq 0$, we deduce that $n = k$. This contradiction shows that $\iota(n) \neq \iota(k)$ and hence ι is injective. Furthermore, for every $n,k \in \mathbb{Z}$ we have

$$\begin{aligned}\iota(n)+\iota(k) &= nu/u + kv/v = ((nu)v + u(kv))/uv = (n(uv)+k(uv))/uv \\ &= (n+k)uv/uv = \iota(n+k) \text{ and} \\ \iota(n)\iota(k) &= \frac{nu}{u}\cdot\frac{kv}{v} = \frac{(nu)(kv)}{uv} = \frac{(nk)(uv)}{uv} = \iota(nk).\end{aligned}$$

It follows that ι induces a bijection between $\mathbb{Z}$ and **Im** ι, which respects the operations of addition and multiplication. Thus ι allows us to identify the integer n with its image $\iota(n) = \frac{nu}{u}$, and we write $\frac{nu}{u} = n$ for each $u \in \mathbb{Z}$, where $u \neq 0$.

Then,

$$\frac{n}{k} = \frac{n\cdot 1\cdot 1}{1\cdot 1\cdot k} = \frac{n1}{1}\cdot\frac{1}{1k} = n\left(\frac{1}{k}\right).$$

We write the reciprocal of $k \in \mathbb{Z}$ as k^{-1}. Clearly, the fraction $\frac{1}{k}$ is the reciprocal to the element $\frac{k1}{1} = k$. So, for the fraction $\frac{n}{k}$, we have

$$\frac{n}{k} = \frac{n1}{1} \cdot \frac{1}{k} = nk^{-1}$$

The existence of reciprocals allows us to define the operation of division on nonzero fractions. We write $\left(\frac{u}{v}\right)^{-1}$ for the reciprocal of $\frac{u}{v}$ and note that $\left(\frac{u}{v}\right)^{-1} = \frac{v}{u}$. We define the quotient of two fractions $\frac{n}{k}$ and $\frac{u}{v}$, where $k, u, v \neq 0$ by

$$\frac{n}{k} \div \frac{u}{v} = \frac{\frac{n}{k}}{\frac{u}{v}} = \frac{n}{k} \cdot \left(\frac{u}{v}\right)^{-1} = \frac{n}{k} \cdot \frac{v}{u} = \frac{nv}{ku}.$$

There are further important properties of rational numbers. For example, we can define an order on the set $\mathbb{Q}$ as follows.

Definition 2.4.5. *Let $\frac{n}{k} \in \mathbb{Q}$. If the integer nk is nonnegative, so $nk \geq 0$, or positive ($nk > 0$), then we say that $\frac{n}{k}$ is nonnegative (respectively positive). In this case, we will write $\frac{n}{k} \geq 0$ (respectively $\frac{n}{k} > 0$). Let $\frac{n}{k}$, $\frac{m}{t} \in \mathbb{Q}$. If $\frac{n}{k} - \frac{m}{t}$ is nonnegative (respectively positive), then we say that "$\frac{n}{k}$ is greater than or equal to $\frac{m}{t}$" or that "$\frac{m}{t}$ is less than or equal to $\frac{n}{k}$" and write these in the former case as $\frac{n}{k} \geq \frac{m}{t}$ and in the latter case as $\frac{m}{t} \leq \frac{n}{k}$. If, in this case, $\frac{n}{k} \neq \frac{m}{t}$, we say that "$\frac{n}{k}$ is greater than $\frac{m}{t}$" or that "$\frac{m}{t}$ is less than $\frac{n}{k}$" and write these in the former case as $\frac{n}{k} > \frac{m}{t}$ and in the latter case as $\frac{m}{t} < \frac{n}{k}$.*

Let $n, k \in \mathbb{Z}$. Consider $\frac{nu}{u} - \frac{kv}{v}$. We have

$$\frac{nu}{u} - \frac{kv}{v} = \frac{(n(uv) - u(kv))}{uv} = \frac{(n(uv) - k(uv))}{uv} = \frac{(n-k)uv}{uv}.$$

Since $(uv)^2 > 0$, if $(n-k)(uv)^2 \geq 0$ (respectively, $(n-k)(uv)^2 > 0$) then $n \geq k$ (respectively, $n > k$). This shows that the order induced on $\mathbf{Im}\ \iota \equiv \mathbb{Z}$ from $\mathbb{Q}$ coincides with the natural order on $\mathbb{Z}$.

It is not hard to see that for every pair of rational numbers, x and y, one and only one of the following relations is valid $x = y$, $x < y$, $x > y$.

As with the set of integers, we can define the absolute value of a rational number, x, by

$$|x| = \begin{cases} x, \text{if } x \geq 0, \\ -x, \text{ if } x < 0 \end{cases}.$$

The following assertions hold for arbitrary $x, y \in \mathbb{Q}$, as can be easily shown:

$|x| \geq 0$, and $|x| = 0$ if and only if $x = 0$;
$|xy| = |x|\,|y|$;
$|x+y| \leq |x| + |y|$.

The rational numbers are those numbers that can be represented as fractions and naturally the question arises as to whether there are numbers that cannot be so represented. A number that cannot be represented as a ratio of two integers is called *irrational*. Real numbers were used by the ancient Greek mathematicians and they already knew that $\sqrt{n}$ need not be a rational number, when $n \in \mathbb{N}$. Indeed, the followers of Pythagoras, the Pythagoreans, discovered that $\sqrt{2}$ is irrational in the fifth century BC. Their discovery actually means that there is no common unit of measure such that the hypotenuse of an isosceles right triangle and one of the legs can both be expressed as a whole number multiple of that unit. At that time, this was a really shocking discovery; the Pythagoreans believed, for example, that all phenomena in the universe could be reduced to whole numbers and their ratios. Here is a simple, elegant proof that $\sqrt{2}$ is irrational. This proof is published in the famous *Elements* of Euclid.

We are going to prove that there is no rational number $r = \frac{p}{q}$, such that $(\frac{p}{q})^2 = 2$. Thus we shall prove that there is no fraction whose square is 2.

To begin the proof, suppose that there exists a rational number r such that $r^2 = 2$. Then there exist integers p and q, with no common divisors, such that $r = \frac{p}{q}$, so the fraction $\frac{p}{q}$ is in its simplest form. Such a fraction is called *irreducible.*

Next square both sides of the equation $\frac{p}{q} = \sqrt{2}$ to obtain $(\frac{p}{q})^2 = 2$. It follows that $2q^2 = p^2$, so p^2 is divisible by 2. Since 2 is a prime number we deduce that 2 also divides p, and hence there is a whole number s such that $p = 2s$. Substituting $p = 2s$ into the equation $2q^2 = p^2$ and dividing both sides by 2, we obtain $q^2 = 2s^2$.

By repeating the argument given here, we deduce that q is also divisible by 2. This means that both p and q are even integers, and the fraction $\frac{p}{q}$ is reducible. Since we initially assumed that $\frac{p}{q}$ is irreducible, we obtain a contradiction. We deduce that there is no rational number r such that $r^2 = 2$ and hence $\sqrt{2}$ is irrational.

However, the number $\sqrt{2}$ clearly exists since, by the Pythagorean Theorem, the square with sides 1 unit long has a diagonal whose length is equal to $\sqrt{2}$ units. Actually it is not difficult, using a method similar to that given above for $\sqrt{2}$, to prove that the square (and other) roots of many natural numbers are irrational. Some other examples of irrational numbers are π, $\ln 10$,

$0.1234567891011\ldots$, $1.01001000100001\ldots$, and many others. The set of rational numbers and the set of irrational numbers together form the set $\mathbb{R}$ of *real numbers*.

For real numbers, the rules of equality, addition, and multiplication are the same as for fractions (see Theorem 2.4.4). The "rule of signs" for multiplication/division is the same as for integers. The real numbers were first constructed axiomatically in the nineteenth century. However, we shall not give a strict construction of the set of real numbers, since this is done in detail in a course concerned with real analysis. We note here two types of real number that are important in algebra courses, namely, transcendental and algebraic numbers.

An *algebraic number* is a number that is a root of a nonzero polynomial equation in one variable with rational coefficients. Numbers that are not algebraic are said to be *transcendental* (so a transcendental number is not a root of any nonzero polynomial equation with rational coefficients). All rational numbers are algebraic. Indeed, if the number r is rational, it is a root of the equation $x - r = 0$ with rational coefficients. Thus all real transcendental numbers must be irrational. The converse is not true however since $\sqrt{2}$ is irrational but not transcendental, since it is a root of $x^2 - 2 = 0$. Transcendental numbers were first shown to exist by Liouville in 1844 and he also gave the first example in 1851. Hermite showed that e is transcendental in 1873. Lindemann demonstrated that π is transcendental in 1882. Using the methods of G. Cantor, one can show that most real and complex numbers are transcendental. By this we mean that the set of algebraic numbers is countable (which informally means that they can be written in a list), while the set of transcendental numbers is uncountable (which informally means they cannot be listed). However, to actually show that a given number is transcendental can be very difficult. A. O. Gelfond built a far-reaching theory, allowing us to determine transcendence for a wide set of numbers that arise naturally in analysis. This theory requires the use of very powerful techniques that are somewhat removed from algebra.

Exercise Set 2.4

In each of the following questions, explain your reasoning by giving either a proof of your assertion or a counterexample.

2.4.1. The natural numbers a, b, and c form a Pythagorean triple if $a^2 + b^2 = c^2$. Here is a method for creating these numbers. Take any odd number, call it a. Square a then subtract 1 and divide by 2 (this gives the second number b). Add 1 to b, and this gives the third number c. Prove this rule.

2.4.2. The natural numbers a, b, and c form a Pythagorean triple if $a^2 + b^2 = c^2$. Show that there are infinitely many Pythagorean triples.

2.4.3. Prove that the infinite decimal $x = 0.12345678...$ is irrational.

2.4.4. Prove that the infinite decimal $x = 0.1001100011100001111...$ is irrational.

2.4.5. Prove that the number $\sqrt{5}$ is irrational.

2.4.6. Prove that for each prime number a, the number $\sqrt{a}$ is irrational.

2.4.7. Prove that the number $\sqrt[3]{3}$ is irrational.

2.4.8. Prove that the number $\sqrt[t]{k}$ is irrational, where t is a natural number greater than 1, and k is a prime.

2.4.9. Prove that the number $\log 3$ is irrational.

2.4.10. Prove that the number $\log t$ is irrational, where t is a prime number.

2.4.11. For the equivalence relation defined in Definition 2.4.1, show in detail that the relation has the transitive property.

2.4.12. If α is rational and β is irrational then are $\alpha + \beta$ and $\alpha\beta$ rational or irrational? Prove your assertion.

2.4.13. Let α be an algebraic number and let r be a rational number. Prove that $r\alpha$ is an algebraic number.

3

GROUPS

3.1 GROUPS AND SUBGROUPS

The first important algebraic structures that were introduced were groups, whose study was motivated by the old problem of finding a formula for the roots of a polynomial in terms of the coefficients of that polynomial. At some stage, work on this problem led mathematicians to study the set of permutations of the roots of a given polynomial. They found that it was crucial to investigate not just this set, but products of these permutations. The first results of this permutation theory, which can be considered as the first results of group theory, were obtained by Lagrange, Vandermonde, Gauss, Ruffini, Cauchy, and Abel.

However, Evariste Galois is recognized as the founder of group theory, since he reduced the study of algebraic equations to the study of permutation groups. He introduced the concept of a normal subgroup and understood its importance. He also considered groups having special given properties and introduced the idea of a linear presentation of a group that is very close to the concept of a homomorphism. His brilliant work was not understood for a long period of time; his ideas were disseminated by J. Serre and C. Jordan, long after his tragic death.

An Introduction to Essential Algebraic Structures, First Edition. Martyn R. Dixon, Leonid A. Kurdachenko and Igor Ya. Subbotin.

Some of the results obtained by Galois initiated the study of groups of permutations. Indeed, at first, group theory was discussed only in terms of groups of permutations. However, fairly soon it was discovered that almost all properties investigated in this theory were indifferent to the specific nature of the elements of the set considered. The definition of an abstract group was introduced by A. Cayley in the middle of the nineteenth century. This transformed group theory into an axiomatic theory, streamlining the results and facilitating its further development.

The heyday of the theory of finite groups lasted from the end of the nineteenth century until the latter part of the twentieth century. During this time, the most important basic results of this theory were proved and the classical methods of research created. The theory of finite groups acquired its own nature and its essential features were discovered, bringing us to the current state of the subject.

For a group to be finite is a very strong restriction and it is not always a natural limitation. During the latter part of the nineteenth century and beginning of the twentieth, different parts of geometry, including the theory of automorphic functions and topology, began coalescing with group theory. Geometers began meeting algebraic objects that turned out to be groups, usually infinite ones. It took quite a long time to understand the relationship between groups and the ideas of invariance and symmetry. Groups appear everywhere where symmetrical properties of an object (algebraic or differential equations, crystal lattices, geometric figures) play a key role. Groups are one measure of the symmetry of an object and that is why they are so important for the classification of such objects. These are the main reasons why groups are of vital importance in different branches of mathematics, physics, chemistry, and so on. The methods of finite group theory were not always adequate for the study of such infinite groups.

Groups with no limitation on finiteness were first considered in the book "Abstract Theory of Groups" (Kiev, 1916) by O. U. Schmidt. However, the extensive development of the general theory of groups began a little later and was associated with the radical restructuring, and the transition to set-theoretic foundations, of algebra made by such people as Emmy Noether in the twentieth century. Thanks to this, new concepts such as systems of operators and chain conditions were introduced into group theory.

To date, group theory has become extremely rich in concepts, results, and applications that have given it a pre-eminent place in modern algebra. Our goal here is not to give a deep exploration of the concepts and methods of group theory. We will give a short introduction, and confine our attention to the most essential concepts and results.

We begin with the standard axiomatic definition of a group.

Definition 3.1.1. *A group is a set G, together with a given binary operation,*

$$(x, y) \longmapsto xy, \text{ where } x, y \in G,$$

satisfying the properties (the group axioms)

(G 1) *the operation is associative, so for all elements $x, y, z \in G$ the equation*

$$x(yz) = (xy)z \text{ holds};$$

(G 2) *G has an identity element, an element $e \in G$ having the property*

$$xe = ex = x$$

for all $x \in G$;

(G 3) *every element $x \in G$ has an inverse, $x^{-1} \in G$, an element such that*

$$xx^{-1} = x^{-1}x = e.$$

We stress that, since the operation is binary, $xy \in G$ whenever $x, y \in G$. This is really a fourth axiom that must be verified when showing that a particular structure is a group.

Definition 3.1.2. *Let G be a group. If the group operation is commutative, then the group is called abelian.*

Often, additive notation is used for abelian groups so, in this case, the group axioms take the following form:

(AG 1) the operation is commutative, so

$$x + y = y + x$$

for all $x, y \in G$;

(AG 2) the operation is associative,

$$x + (y + z) = (x + y) + z \text{ for all } x, y, z \in G;$$

(AG 3) G has a zero element, an element $0_G \in G$ having the property that

$$x + 0_G = 0_G + x = x$$

for all $x \in G$;

(AG 4) every element $x \in G$ has a negative element, $-x \in G$, an element such that

$$x + (-x) = (-x) + x = 0_G.$$

We note here that 0_G is not necessarily the integer 0.

We next give some **examples** of groups.

1. The set $\mathbb{Z}$ of all integers is easily seen to be an abelian group using the binary operation of addition. It is well-known that this operation is commutative and associative. In this case the number 0 is the zero element of $\mathbb{Z}$, since $0 + a = a + 0 = a$ for all $a \in \mathbb{Z}$. Furthermore, for each $a \in \mathbb{Z}$ the number $-a$ is an integer and $-a + a = a + (-a) = 0$. Since $\mathbb{Z}$ is an infinite set, this group is infinite.

   However, the set $\mathbb{Z}$ does not form a group under the operation of multiplication. Certainly, multiplication is commutative and associative. Furthermore the number 1 is the (multiplicative) identity element since $1 \cdot a = a \cdot 1 = a$, for all $a \in \mathbb{Z}$, but only two elements, 1 and -1 have multiplicative inverses that belong to the set $\mathbb{Z}$. It is essential that the inverse of an element actually belongs to the set under consideration.
2. Similarly, the set $\mathbb{Q}$ of all rational numbers and the set $\mathbb{R}$ of all real numbers form abelian groups under addition. Of course, these groups are also infinite.

   The set $\mathbb{Q}^* = \mathbb{Q} \setminus \{0\}$ of nonzero rationals is an abelian group under multiplication. Again, it is well-known that this operation is commutative and associative. In this case the (multiplicative) identity is the number 1, since $1 \cdot a = a = a \cdot 1$, for all $a \in \mathbb{Q}^*$. Furthermore in $\mathbb{Q}^*$ every element contains a multiplicative inverse which is also a rational number, namely the reciprocal of the number. This group is infinite since the set $\mathbb{Q}^*$ is infinite.

   In the same way the set $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$ of nonzero real numbers is an infinite abelian group under multiplication.
3. In Section 1.3 we considered the set $\mathbf{S}_n$ of all permutations of degree n. We recall that multiplication of permutations, namely composition of mappings, is associative. The identity element is the permutation $\varepsilon = \begin{pmatrix} 1 & 2 & 3 & \ldots & n \\ 1 & 2 & 3 & \ldots & n \end{pmatrix}$, and every permutation has an inverse, since a permutation in $\mathbf{S}_n$ is a bijection of the set $\{1, 2, \ldots, n\}$. Thus it is easy to see, from our prior knowledge, that the set $\mathbf{S}_n$ of all permutations of degree n is a group, with this multiplication. As we saw earlier, multiplication of

permutations is not commutative, so that the group $\mathbf{S}_n$ is nonabelian. The group is finite—we proved in Section 1.3 that $|\mathbf{S}_n| = n!$.

4. In Section 1.4 we considered the set $\mathbf{M}_n(\mathbb{R})$ of all $n \times n$ matrices whose entries belong to the set $\mathbb{R}$ of all real numbers.

   Two binary operations were defined on this set, matrix addition and matrix multiplication. We saw that matrix addition is commutative and associative. The set $\mathbf{M}_n(\mathbb{R})$ has a zero element, the matrix O, all of whose entries are 0, and every matrix $A = (a_{ij})$ has an additive inverse, namely $-A = (-a_{ij})$. Thus we see that $\mathbf{M}_n(\mathbb{R})$ is an abelian group under addition. The situation with multiplication of matrices is more complicated. As we saw in Section 1.4, multiplication of matrices is associative, and there is a multiplicative identity element, namely the $n \times n$ identity matrix I. However not every matrix has a multiplicative inverse. In Section 1.4, we exhibited a nonzero matrix that has no inverse. However, the set $\mathbf{GL}_n(\mathbb{R})$ of all nonsingular $n \times n$ matrices is a group since every matrix in this set has an inverse that is also nonsingular. In Section 1.4 we found two nonsingular matrices A, B such that $AB \neq BA$. This means that $\mathbf{GL}_n(\mathbb{R})$ is a nonabelian group. Furthermore, this group is infinite. The group $\mathbf{GL}_n(\mathbb{R})$ is called the General Linear group of order (or degree) n with real coefficients.

5. Let $E = \mathbb{R}^n$. If $x, y \in E$, then let $d(x,y)$ denote the distance between the points x and y. A bijective mapping $f : E \longrightarrow E$ is called an *isometry* of the space E if $d(f(x),f(y)) = d(x,y)$ for all $x, y \in E$. We denote the set of all isometries of E by $\mathbf{Isom}(E)$. If $f, g \in \mathbf{Isom}(E)$, $x, y \in E$, then

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

   so $f \circ g \in \mathbf{Isom}(E)$. Hence $\mathbf{Isom}(E)$ is a closed subset of the set of all bijective mappings $P(E)$, so that

$$(f,g) \longmapsto f \circ g,\ f, g \in \mathbf{Isom}(E),$$

   is a binary operation on $\mathbf{Isom}(E)$. The operation is associative, because multiplication of mappings is associative. The map ε_E defined by $\varepsilon_E(x) = x$ for all $x \in E$ is clearly an element of $\mathbf{Isom}(E)$, thus ε_E is the identity element of $\mathbf{Isom}(E)$. Since every isometry f is bijective, Corollary 1.3.5 shows that f^{-1} exists. Let $x, y \in E$, then

$$d(x,y) = d(f(f^{-1}(x)), f(f^{-1}(y))) = d(f^{-1}(x), f^{-1}(y)),$$

   so $f^{-1} \in \mathbf{Isom}(E)$ also. Consequently, the conditions **(G 1)–(G 3)** are satisfied, which shows that $\mathbf{Isom}(E)$ is a group under composition of mappings.

6. Consider the set $D = \{0, 1, 2, 3, 4\}$. Define the operation $\boxplus$ on D by the rule:

$$k \boxplus n = \begin{cases} k+n, \text{ if } k+n < 5 \\ m, \text{ where } m \text{ is the remainder when } k+n \geq 5 \text{ is divided by } 5, \end{cases}$$

Since $k+n = n+k$ (and $k+(n+t) = (k+n)+t$), the remainders upon division of $k+n$ and $n+k$ (respectively $k+(n+t)$ and $(k+n)+t$) by 5 coincide. It follows that the operation $\boxplus$ is commutative and associative. Clearly 0 is the zero element for this operation. Every number in D has an additive inverse in D. In fact, for a number $k \in D$ this inverse is $5-k \in D$. This shows that D is an abelian group under the operation $\boxplus$.

Now we will introduce a very important concept, namely that of a subgroup.

Definition 3.1.3. *A closed subset H of a group G is called a subgroup of G, if H is itself a group under the operation defined on G. The fact that H is a subgroup of G will be denoted by $H \leq G$.*

Clearly, for each group G, the subset $\{e\}$ (or $\{0_G\}$, if the operation is addition) is a subgroup of G. Notice also that every group is a subgroup of itself. A subgroup A of a group G is said to be *proper* if $A \neq G$.

It is usually inconvenient to verify the group axioms for a subgroup. The following theorem gives a useful shortcut for showing that a given subset is a subgroup.

Theorem 3.1.4. *(Subgroup Criterion). Let G be a group. If H is a subgroup of G, then H satisfies the conditions:*

(SG 1) *if $x, y \in H$, then $xy \in H$;*
(SG 2) *if $x \in H$, then $x^{-1} \in H$.*

*Conversely, if H is a nonempty subset of G satisfying the conditions **(SG 1)** and **(SG 2)**, then H is a subgroup of G.*

Proof. If H is a subgroup, then the condition **(SG 1)** restates the fact that the operation is closed on H. Let e_H be the identity element of H. Then for an arbitrary element $x \in H$ we have $xe_H = x$. Furthermore, there exists an element $y \in H$ such that $xy = yx = e_H$. Since G is a group, then x has an inverse in the whole of G, so there exists $x^{-1} \in G$ such that $xx^{-1} = x^{-1}x = e_G$. Then from $xe_H = x$ we obtain $x^{-1}(xe_H) = x^{-1}x = e_G$. Since $x^{-1}(xe_H) = (x^{-1}x)e_H = e_G e_H = e_H$, we have $e_H = e_G$. Thus, each subgroup contains the identity element e_G of the entire group G. Then $xy = e_G = xx^{-1}$ and $y = x^{-1}$, as we showed in Section 1.5. Hence (**SG 2**) is also valid.

Conversely, let H be a nonempty subset of G satisfying the conditions (**SG 1**) and (**SG 2**). From condition (**SG 1**) we deduce that H is a closed subset of G. It follows that the operation defined on G, restricted to H, is a binary operation on H. This operation is associative, since the original operation on the entire group G is associative. If $x \in H$, then (**SG 2**) implies that $x^{-1} \in H$. Then by (**SG 1**), $e_G = xx^{-1} \in H$, so that e_G is the identity element of H. Finally, condition (**G 3**) of Definition 3.1.1 follows from (**SG 2**).

The number of conditions that must be checked to see that a particular subset is a subgroup can be reduced still further, as follows. It is important to realize that we must also check that $H \neq \emptyset$.

Corollary 3.1.5. *Let G be a group. If H is a subgroup of G, then H satisfies the condition:*

(SG 3) *if $x, y \in H$, then $xy^{-1} \in H$.*

*Conversely, if H is a nonempty subset of G satisfying the condition **(SG 3)**, then H is a subgroup of G.*

Proof. We will show that the condition **(SG 3)** is equivalent to conditions **(SG 1)** and **(SG 2)**. Let H be a subgroup of G and let $x, y \in H$. Then by Theorem 3.1.4, $y^{-1} \in H$. Using condition (**SG 1**) we deduce that $xy^{-1} \in H$, so H satisfies (**SG 3**).

Conversely, let H be a nonempty subset of G satisfying (**SG 3**). Then H contains at least one element x and using condition (**SG 3**), we obtain $e = xx^{-1} \in H$.

Furthermore, an application of (**SG 3**) gives $x^{-1} = ex^{-1} \in H$, so that **(SG 2)** holds. Finally, let $y \in H$. We have already proved that, in this case, $y^{-1} \in H$. Therefore $xy = x(y^{-1})^{-1} \in H$, again using (**SG 3**) and the fact that $(y^{-1})^{-1} = y$. Hence **(SG 1)** holds.

Since Condition **(SG3)** is dependent only on the subset in question the following result, that a subgroup of a subgroup of a group is also a subgroup of the group, is immediate.

Corollary 3.1.6. *Let G be a group and let H be a subgroup of G. A subset K of H is a subgroup of G if and only if K is a subgroup of H.*

The additive version of Corollary 3.1.5 can be easily obtained by noting that if G is an additive abelian group then the operation of subtraction on G can be defined by $x - y = x + (-y)$. Then, Corollary 3.1.5 is as follows.

Corollary 3.1.7. *(Additive form). Let G be an abelian group under addition. If H is a subgroup of G, then H satisfies the condition:*

(ASG 3) *if $x, y \in H$, then $x - y \in H$.*

*Conversely, if H is a nonempty subset of G satisfying the condition (**ASG 3**), then H is a subgroup of G.*

Here are some examples of subgroups.

1. As we mentioned above, the set of integers forms an abelian group under addition. The nonempty subset $2\mathbb{Z}$, consisting of all even integers, the set of multiples of 2, is a subgroup of $\mathbb{Z}$. Indeed, let $x, y \in 2\mathbb{Z}$, so that $x = 2k, y = 2m$ for certain integers k, m. Then $x - y = 2k - 2m = 2(k - m) \in 2\mathbb{Z}$, since $k - m \in \mathbb{Z}$.

   Thus $2\mathbb{Z}$ satisfies (**ASG 3**) and therefore is a subgroup of $\mathbb{Z}$. Essentially the same argument shows that the subset $n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}$ is a subgroup of $\mathbb{Z}$ for every positive integer n. On the other hand, the subset of all odd integers is not a subgroup since the sum of two odd integers is even.
2. Next we consider some subgroups of the additive group $\mathbb{Q}$ of rational numbers. Clearly $\mathbb{Z}$ is a subgroup of $\mathbb{Q}$. Let

$$\mathbb{Q}_3 = \left\{ \frac{m}{3^k} \mid m, k \in \mathbb{Z} \right\}.$$

   Since

$$\frac{m}{3^k} - \frac{r}{3^s} = \frac{m3^s - r3^k}{3^{k+s}},$$

   for $m, r, k, s \in \mathbb{Z}$, it is easy to see that $\mathbb{Q}_3$ satisfies condition (**ASG 3**). Hence $\mathbb{Q}_3$ is a subgroup of $\mathbb{Q}$. In a similar way, if p is a prime, then the subset

$$\mathbb{Q}_p = \left\{ \frac{m}{p^k} \mid m, k \in \mathbb{Z} \right\}.$$

   is also a subgroup of $\mathbb{Q}$. Thus $\mathbb{Q}$ has infinitely many proper subgroups.
3. As we noted above, the set $\mathbb{Q}\backslash\{0\} = \mathbb{Q}^*$ of all nonzero rationals is an abelian group under multiplication. The subset H of all positive rationals is a subgroup of $\mathbb{Q}^*$. Clearly, if $x, y \in \mathbb{Q}$ and $x, y > 0$, then $xy^{-1} > 0$ and Corollary 3.1.5 shows that H is a subgroup of $\mathbb{Q}^*$. It is easy to see that $\{1, -1\}$ is also a subgroup of $\mathbb{Q}^*$. We note that $\mathbb{Q}^*$ has infinitely many proper subgroups. For example, consider the nonempty subset $\{5^n \mid n \in \mathbb{Z}\}$. We have

$5^n(5^k)^{-1} = 5^n 5^{-k} = 5^{n-k}$ and Corollary 3.1.5 shows that this subset is a subgroup. By the same argument, the subset $\{p^n \mid n \in \mathbb{Z}\}$ is easily seen to be a proper subgroup of $\mathbb{Q}^*$, for each prime p.

4. As we saw above the set $\mathbb{R}\backslash\{0\} = \mathbb{R}^*$ of all nonzero real numbers is an abelian group under multiplication, containing the subgroup $\mathbb{Q}^*$. It is easy to see that the subset $\mathbb{R}^+$ of all positive real numbers is also a subgroup of $\mathbb{R}^*$
5. The set $\mathbb{C}\backslash\{0\} = \mathbb{C}^*$ of all nonzero complex numbers is an abelian group under multiplication and contains the subgroups $\mathbb{R}^*$ and $\mathbb{Q}^*$.

   Consider the nonempty subset

$$\mathbb{T} = \{\alpha \in \mathbb{C} \mid \alpha^n = 1 \text{ for some positive integer } n\},$$

   the set of complex roots of unity. Let $\alpha, \beta \in \mathbb{T}$ and n, k be positive integers such that $\alpha^n = 1 = \beta^k$. Then $(\alpha\beta)^{nk} = 1$, so $\mathbb{T}$ satisfies condition (**SG 1**). Clearly

$$\left(\frac{1}{\alpha}\right)^n = \frac{1}{\alpha^n} = 1,$$

   which shows that $\alpha^{-1} = \frac{1}{\alpha} \in \mathbb{T}$. Thus $\mathbb{T}$ satisfies condition (**SG 2**), and hence $\mathbb{T}$ is a subgroup of $\mathbb{C}^*$

   Furthermore, let $\mathbb{T}_1 = \{\alpha \in \mathbb{C} \mid |\alpha| = 1\}$, a nonempty subset. If α, β are arbitrary elements of $\mathbb{T}_1$, then $\left|\alpha\beta^{-1}\right| = \left|\alpha\right|\left|\beta^{-1}\right| = 1 \cdot 1 = 1$. Thus $\mathbb{T}_1$ satisfies the condition (**SG 3**) and hence $\mathbb{T}_1$ is a subgroup of $\mathbb{C}^*$. The subgroup $\mathbb{T}_1$ is called a *one-dimensional torus*, in this case the unit circle. Clearly $\mathbb{T}$ is a subgroup of $\mathbb{T}_1$.

   Next let k be a fixed positive integer, and let $\mathbb{C}_k = \{\alpha \in \mathbb{C} \mid \alpha^k = 1\}$, a nonempty subset. If $\alpha, \beta \in \mathbb{C}_k$ then we have $(\alpha\beta)^k = \alpha^k\beta^k = 1$, so that $\mathbb{C}_k$ satisfies the condition (**SG 1**). Clearly also,

$$(\alpha^{-1})^k = \alpha^{-k} = \left(\frac{1}{\alpha^k}\right) = 1,$$

   which shows that $\mathbb{C}_k$ satisfies condition (**SG 2**) and hence $\mathbb{C}_k$ is a subgroup of $\mathbb{C}^*$. We note that, in particular, $\mathbb{C}_2 = \{1, -1\}, \mathbb{C}_4 = \{1, -1, i, -i\}$, and so on.
6. We next consider some subgroups of the group $\mathbf{S}_n$ of all permutations of degree n. Recall that the product of two even permutations is even. Also the inverse of an even permutation is likewise even. Thus the nonempty

subset $\mathbf{A}_n$ of all even permutations satisfies the conditions (**SG 1**), (**SG 2**) and hence is a subgroup of $\mathbf{S}_n$. Further, for every transposition ι_{kt} we have

$$\iota_{kt}(\iota_{kt}(k)) = \iota_{kt}(t) = k, \text{ and } \iota_{kt}(\iota_{kt}(t)) = \iota_{kt}(k) = t.$$

Also, if $j \notin \{k,t\}$, then $\iota_{kt}(\iota_{kt}(j)) = \iota_{kt}(j) = j$.

Thus $\iota_{kt}^2 = \varepsilon$. It follows that the subset $\{\iota_{kt}, \varepsilon\}$ satisfies the conditions (**SG 1**), (**SG 2**), and hence is a subgroup of $\mathbf{S}_n$ for each transposition ι_{kt}.

If a group G is finite, then the way that the operation is defined can be given in the form of a table. In the first row and the first column we write down all the elements of the group, and at the intersection of the row corresponding to the element x and the column corresponding to y, we write the element xy. This table is called the *Cayley table* of the group.

The Cayley table for the group $\mathbf{S}_3$ can be completed as follows. Let

$$\iota_{12} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \iota_{13} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \iota_{23} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}$$

be the transpositions.

The other nonidentity permutations are $\lambda = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$ and $\mu = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}$.

When all pairs of permutations are multiplied, we obtain the following table (recall that for permutations the product $\pi \circ \sigma$ is defined so that π is done first).

$\circ$	ε	ι_{12}	ι_{13}	ι_{23}	λ	μ
ε	ε	ι_{12}	ι_{13}	ι_{23}	λ	μ
ι_{12}	ι_{12}	ε	μ	λ	ι_{23}	ι_{13}
ι_{13}	ι_{13}	λ	ε	μ	ι_{12}	ι_{23}
ι_{23}	ι_{23}	μ	λ	ε	ι_{13}	ι_{12}
λ	λ	ι_{13}	ι_{23}	ι_{12}	μ	ε
μ	μ	ι_{23}	ι_{12}	ι_{13}	ε	λ

We shall obtain all the subgroups of $\mathbf{S}_3$. Always there are two easy subgroups: $\{\varepsilon\}$ and $\mathbf{S}_3$. As done earlier, we obtain three subgroups $T_1 = \{\iota_{12}, \varepsilon\}$, $T_2 = \{\iota_{13}, \varepsilon\}, T_3 = \{\iota_{23}, \varepsilon\}$. Furthermore, $\lambda^2 = \mu, \lambda^3 = \varepsilon$, and hence $\lambda^2 = \lambda^{-1} = \mu, \mu^2 = \lambda^4 = \lambda^3 \circ \lambda = \varepsilon \circ \lambda = \lambda, \mu^3 = \lambda^6 = \lambda^3 \circ \lambda^3 = \varepsilon \circ \varepsilon = \varepsilon$. These equations show that the subset $L = \{\lambda, \mu, \varepsilon\}$ satisfies the conditions (**SG 1**), (**SG 2**), and hence it is a subgroup of $\mathbf{S}_3$. Let H be another subgroup of $\mathbf{S}_3$. The Cayley table of $\mathbf{S}_3$ shows that if H contains two different transpositions,

then $H = G$. If H contains λ and one transposition, then $H = G$. Similarly, if H contains μ and one transposition, then $H = G$. This shows that G contains no other subgroups from those mentioned here.

Next we consider $\mathbf{S}_6$ and obtain some (but not all) of its subgroups. First let

$$K_{2,4} = \{\pi \in \mathbf{S}_6 \mid \pi(2) = 2, \pi(4) = 4\},$$

which is a nonempty set. If $\lambda, \pi \in K_{2,4}$, then

$$(\lambda \circ \pi)(2) = \pi(\lambda(2)) = \pi(2) = 2, (\lambda \circ \pi)(4) = \pi(\lambda(4)) = \pi(4) = 4,$$

which shows that $K_{2,4}$ satisfies condition (**SG 1**). Also, when $\lambda \in K_{2,4}$ then we have

$$2 = \varepsilon(2) = (\lambda \circ \lambda^{-1})(2) = (\lambda^{-1}(\lambda(2)) = \lambda^{-1}(2),$$

and likewise, $\lambda^{-1}(4) = 4$. Thus $K_{2,4}$ satisfies condition (**SG 2**), and hence is a subgroup $\mathbf{S}_6$.

Next we consider a slightly different case. Let

$$L_{2,4} = \{\pi \in \mathbf{S}_6 \mid \{\pi(2), \pi(4)\} = \{2, 4\}\}.$$

If $\lambda, \pi \in L_{2,4}$, and $j \in \{2, 4\}$, then $(\pi \circ \lambda)(j) = \lambda(\pi(j))$. Since $\pi \in L_{2,4}$, $\pi(j) = k \in \{2, 4\}$ and since $\lambda \in L_{2,4}$ we have $\lambda(k) \in \{2, 4\}$, so that $(\pi \circ \lambda)(j) \in \{2, 4\}$. Hence $\pi \circ \lambda \in L_{2,4}$, which shows that $L_{2,4}$ satisfies condition (**SG 1**). Also if $j \in \{2, 4\}$ and $\pi \in L_{2,4}$ then there exists $k \in \{2, 4\}$ such that $j = \pi(k)$. We now have

$$k = \varepsilon(k) = (\pi \circ \pi^{-1})(k) = \pi^{-1}(\pi(k)) = \pi^{-1}(j),$$

so $\pi^{-1} \in L_{2,4}$. Consequently, $L_{2,4}$ satisfies condition (**SG 2**), and hence is a subgroup of $\mathbf{S}_6$. Note also that $K_{2,4} \leq L_{2,4}$. Clearly we can form, similarly, the natural subgroups $L_{i,j}$ and $K_{i,j}$, for $i \neq j$, in $\mathbf{S}_n$ in general.

7. We have already shown that the subset $\mathbf{GL}_n(\mathbb{R})$ of all nonsingular matrices in the set $\mathbf{M}_n(\mathbb{R})$ is a group under multiplication. Note that $\mathbf{GL}_n(\mathbb{R})$ is not a subgroup of $\mathbf{M}_n(\mathbb{R})$ since the operations that make these sets into groups are different.

   In a course on linear algebra it is proved that

$$\mathbf{det}(AB) = \mathbf{det}(A)\mathbf{det}(B),$$

for all matrices $A, B \in \mathbf{M}_n(\mathbb{R})$, where $\mathbf{det}(A)$ denotes the determinant of A. Let

$$\mathbf{SL}_n(\mathbb{R}) = \{A \in \mathbf{M}_n(\mathbb{R}) \mid \mathbf{det}(A) = 1\},$$

a nonempty set. If $A, B \in \mathbf{SL}_n(\mathbb{R})$, then $\mathbf{det}(AB) = \mathbf{det}(A)\mathbf{det}(B) = 1$, from which it follows that $AB \in \mathbf{SL}_n(\mathbb{R})$. Furthermore,

$$1 = \mathbf{det}(I) = \mathbf{det}(AA^{-1}) = \mathbf{det}(A)\mathbf{det}(A^{-1}),$$

and therefore $\mathbf{det}(A^{-1}) = 1$, so that $A^{-1} \in \mathbf{SL}_n(\mathbb{R})$ also. Consequently, $\mathbf{SL}_n(\mathbb{R})$ satisfies both conditions (**SG 1**) and (**SG 2**), and hence it is a subgroup of $\mathbf{GL}_n(\mathbb{R})$. This subgroup is called the Special Linear group.

Let $\mathbf{T}_n^0(\mathbb{R})$ denote the subset of all nonsingular upper triangular matrices in the set $\mathbf{M}_n(\mathbb{R})$. We next prove that $\mathbf{T}_n^0(\mathbb{R})$ is a subgroup of $\mathbf{GL}_n(\mathbb{R})$, contenting ourselves here by just considering the case of 3×3 matrices. Let $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix}$, and $B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ 0 & b_{22} & b_{23} \\ 0 & 0 & b_{33} \end{pmatrix}$ be two nonsingular upper triangular matrices, so all the diagonal entries are nonzero. We have

$$AB = \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} + a_{12}b_{22} & a_{11}b_{13} + a_{12}b_{23} + a_{13}b_{33} \\ 0 & a_{22}b_{22} & a_{22}b_{23} + a_{23}b_{33} \\ 0 & 0 & a_{33}b_{33} \end{pmatrix},$$

so that $AB \in \mathbf{T}_3^0(\mathbb{R})$. This shows that $\mathbf{T}_3^0(\mathbb{R})$ satisfies condition (**SG 1**). Next let $A \in \mathbf{T}_3^0(\mathbb{R})$ and let $\mathrm{A}^{-1} = (x_{jk})_{1 \le j,k \le 3}$. We think of the coefficients x_{jk} as variables and use the matrix equation $A^{-1}A = I$ to determine them. We have

$$\begin{aligned} A^{-1}A &= \begin{pmatrix} x_{11}a_{11} & x_{11}a_{12} + x_{12}a_{22} & x_{11}a_{13} + x_{12}a_{23} + x_{13}a_{33} \\ x_{21}a_{11} & x_{21}a_{12} + x_{22}a_{22} & x_{21}a_{13} + x_{22}a_{23} + x_{23}a_{33} \\ x_{31}a_{11} & x_{31}a_{12} + x_{32}a_{22} & x_{31}a_{13} + x_{32}a_{23} + x_{33}a_{33} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \end{aligned}$$

Since A is nonsingular, we have $\mathbf{det}(A) = a_{11}a_{22}a_{33} \neq 0$, so that $a_{11} \neq 0, a_{22} \neq 0, a_{33} \neq 0$. Then from $x_{21}a_{11} = 0 = x_{31}a_{11}$ we deduce that $x_{21} = 0 = x_{31}$. It follows that $x_{32}a_{22} = 0$, which implies that $x_{32} = 0$. Hence we see that the matrix A^{-1} is upper triangular. Consequently, $\mathbf{T}_3^0(\mathbb{R})$ satisfies both conditions (**SG 1**), (**SG 2**), and hence is a subgroup of $\mathbf{GL}_3(\mathbb{R})$.

We note the following important property of subgroups. We recall that if $\mathfrak{S}$ is a family of sets then $\bigcap \mathfrak{S} = \bigcap \{T | T \in \mathfrak{S}\}$

Proposition 3.1.8. *Let G be a group and let $\mathfrak{S}$ be a family of subgroups of G. Then the intersection, $\bigcap \mathfrak{S}$, of the subgroups of this family is a subgroup of G.*

Proof. We note that $T = \bigcap \mathfrak{S}$ is nonempty since, for each subgroup $U \in \mathfrak{S}$, we have $e \in U$ and hence $e \in T$. Let $x, y \in T$ and let U be an arbitrary subgroup of the family $\mathfrak{S}$. Then $xy^{-1} \in U$, by **(SG 3)**. Since this is true for each $U \in \mathfrak{S}$ we have $xy^{-1} \in T$. Corollary 3.1.5 gives the result.

We next discuss a very important method for constructing subgroups.

Let x be an element of a group G and consider the nonempty subset $X = \{x^n \mid n \in \mathbb{Z}\}$. From Proposition 1.5.8 we see that $x^n x^{-m} = x^{n-m}$, so that X satisfies condition (**SG 3**), and hence it is a subgroup of G.

Definition 3.1.9. *The subgroup $\{x^n \mid n \in \mathbb{Z}\} = \langle x \rangle$ is called the cyclic subgroup generated by x. An element y with the property that $\langle y \rangle = \langle x \rangle$ is called a generator of $\langle x \rangle$. A group G is called cyclic if it coincides with at least one of its cyclic subgroups.*

In general several different elements may generate a cyclic group. We now consider the subgroup $\langle x \rangle$ further. There are two cases.

(i) $x^n \neq x^m$, whenever $n \neq m$.
(ii) There exist integers n, m, such that $n \neq m$ but $x^n = x^m$.

In case (i), $\langle x \rangle = \{x^n | n \in \mathbb{Z}\}$ is infinite and x is said to have infinite order. In this case all the integer powers of x are distinct. In case (ii) one of the integers n and m is greater than the other, say $n > m$. From the equation $x^n = x^m$ we deduce that $x^{n-m} = e$, which means that some positive power of x is the identity element. Let $\mathcal{S} = \{k \in \mathbb{N} \mid x^k = e\}$, a nonempty set. Then $\mathcal{S}$ has a least element t. Thus, t is the smallest positive integer such that $x^t = e$. Let n be an arbitrary integer. Then, by Theorem 2.2.1, $n = tq + r$ where $0 \leq r < t$. We have

$$x^n = x^{tq+r} = x^{tq} x^r = (x^t)^q x^r = x^r.$$

By the definition of t the elements

$$e = x^0, x = x^1, x^2, \ldots, x^{t-1}$$

are distinct and it follows that

$$\langle x\rangle = \{x^n \mid n \in \mathbb{Z}\} = \{x^0 = e, x = x^1, x^2, x^{t-1}\}.$$

In this case, we say that the element x has finite order. We record this information in the following important definition.

Definition 3.1.10. *Let G be a group and let $x \in G$. The order of the element x is the least positive integer t, if such exists, such that $x^t = e$, and in this case x is said to have finite order t. If there is no such integer t then x is said to have infinite order. We denote the order of x by $|x|$.*

Thus if x has order t then we write $|x| = t$. Also, by definition, $|x| = |\langle x\rangle|$. We note also that $|e| = 1$.

The next simple but useful result follows from Corollary 3.1.5.

Proposition 3.1.11. *Let G be a group and let H be a subgroup of G. If $x \in H$, then $\langle x\rangle \leq H$.*

We next show the important fact that a subgroup of a cyclic group is also cyclic.

Theorem 3.1.12. *Let $G = \langle g\rangle$ be a cyclic group. If H is a subgroup of G, then H is also cyclic.*

Proof. Since H is a subgroup, $e \in H$. If $H = \{e\}$, then $H = \langle e\rangle$ is cyclic. Suppose now that H contains nontrivial elements. All elements of G are powers of g. Therefore there exists an integer $m \neq 0$ such that $e \neq g^m \in H$. If $m < 0$, then by the condition **(SG 2)** we have $(g^m)^{-1} = g^{-m} \in H$. This means that H contains positive powers of g that are nontrivial. Let

$$\Omega = \{k > 0 \mid g^k \in H\}.$$

We let d be the least natural number in Ω. In particular, $g^d \in H$ and, by Proposition 3.1.11, $\langle g^d\rangle \leq H$. In fact $H = \langle g^d\rangle$. To see this let x be an arbitrary element of H. Then $x = g^n$ for some $n \in \mathbb{Z}$. By Theorem 2.2.1, $n = dq + r$, where $0 \leq r < d$ and it follows that

$$x = g^n = g^{dq+r} = (g^d)^q(g^r) \text{ and } g^r = x(g^{dq})^{-1}.$$

By **(SG 3),** $g^r \in H$ and therefore $r \in \Omega$ if $r \neq 0$, which contradicts the choice of d. Thus $r = 0$ and hence $x = g^n = (g^d)^q$. This implies that $H \leq \langle g^d\rangle$ and, since $H \geq \langle g^d\rangle$, we have $H = \langle g^d\rangle$, a cyclic group.

Exercise Set 3.1

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

3.1.1. Let $G = \{a + bi\sqrt{5} \mid a, b \in \mathbb{Q}, a^2 + 5b^2 \neq 0\}$. Is G a group under multiplication of complex numbers?

3.1.2. Let $G = \{a + bi\sqrt{3} \mid a, b \in \mathbb{Q}, a^2 + 3b^2 \neq 0\}$. Is G a group under multiplication of complex numbers?

3.1.3. On a set of four elements define a commutative and associative operation having an identity element.

3.1.4. Let $M = \{x, y, z\}$. Define a binary operation such that M is a group under this operation.

3.1.5. Is the set of complex numbers $G = \{\alpha \in \mathbb{C} \mid |\alpha| = 1\}$, a subgroup of the multiplicative group of non-zero complex numbers $\mathbf{U}(\mathbb{C})$?

3.1.6. Is the set of complex numbers $G = \{\alpha \in \mathbb{C} \mid 0 \neq |\alpha| = r\}$ a subgroup of $\mathbf{U}(\mathbb{C})$?

3.1.7. Suppose that $g^2 = e$ for all elements g of a group G. Prove that G is abelian.

3.1.8. Let G be an abelian group. Prove that the subset of all elements of G having finite order is a subgroup.

3.1.9. On the set $G = \mathbb{Z} \times \{-1, 1\}$ we define the operation $\star$ by $(m, a) \star (n, b) = (m + an, ab)$. Is G a group? Is this operation commutative?

3.1.10. On the set $G = \mathbb{Z} \times \{-1, 1\}$ we define the operation $\circ$ by the rule $(m, a) \circ (n, b) = (m + n, ab)$. Is G a group? Is this operation commutative?

3.1.11. Find all the generators of $\mathbb{Z}$.

3.1.12. Let $G = \langle a \rangle$ be a finite cyclic group of order n. Prove that the order of a^k is $n/\mathbf{GCD}(n, k)$ for each integer k.

3.1.13. Let $G = \langle x \rangle$ be a finite cyclic group of order n. Find all the generators of G in the case when (a) $n = 24$, (b) $n = 7$, (c) $n = 18$.

3.1.14. Let $G = \langle x \rangle$ be a cyclic group of order n. Find the order of each of the elements in G in the cases when $n = 12$ and $n = 16$.

3.1.15. Let A be the set of all real numbers excluding -1. Define the operation $\circ$ on A by $a \circ b = a + b + ab$ for all $a, b \in A$. Prove that A together with this operation forms an abelian group.

3.1.16. Let n be a natural number and let $X(n) = \{a \in \mathbb{Z} | 1 \leq a < n$ and $\mathbf{GCD}(a,n) = 1\}$. Write $a \equiv b \pmod{n}$ to mean that n divides $b-a$ (see Exercise 1.5.13). Define a product on $X(n)$ as follows. If $a,b \in X(n)$ then first write $ab = qn+r$, where $0 \leq r < n$, using the division algorithm, and then define $ab = r \in X(n)$. Prove that this product is well-defined and that $X(n)$ is an abelian group under multiplication.

3.1.17. Write the Cayley tables for $X(7)$ and $X(8)$ and determine the order of each element occurring. Are either of these groups cyclic?

3.1.18. Let H,K be subgroups of a group G. Let $HK = \{hk | h \in H, k \in K\}$. Prove that HK is a subgroup of G if and only if $HK = KH$.

3.1.19. Find the order of the element $\begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$ in the group $\mathbf{GL}_2(\mathbb{R})$.

3.1.20. Find the orders of the elements $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and $\begin{pmatrix} 1 & -1 \\ 0 & -1 \end{pmatrix}$ in the group $\mathbf{GL}_2(\mathbb{R})$. Then find the order of their product and explain why this is a bit surprising.

3.2 COSETS AND NORMAL SUBGROUPS

We start this section with some examples. Consider the additive group $\mathbb{Z}$. Let m be a positive integer. In Section 1.5 we defined an equivalence relation on $\mathbb{Z}$ by saying that two integers a,b are congruent modulo m if $a-b$ is divisible by m. We denote this by $a \equiv b \pmod{m}$. Let n be an arbitrary integer. By Theorem 2.2.1, there exist integers q,r such that $n = mq+r$ and $0 \leq r < m$. We write $r+m\mathbb{Z} = \{r+mk \mid k \in \mathbb{Z}\}$. Then $n \in r+m\mathbb{Z}$, and since n is an arbitrary integer, $\mathbb{Z} = \bigcup_{0 \leq r < m} (r+m\mathbb{Z})$. If t is an integer with the property that $t \equiv n \pmod{m}$, then $t-n = ms$ for some integer s, so that

$$t = ms+n = ms+mq+r = m(s+q)+r,$$

which shows that $t \in r+m\mathbb{Z}$. Conversely, if $u \in r+m\mathbb{Z}$, then $u-r \in m\mathbb{Z}$, so $u \equiv r \pmod{m}$. This shows that $r+m\mathbb{Z}$ is exactly the equivalence class of u, under this equivalence relation. In particular there are m equivalence classes, $0+m\mathbb{Z}, 1+m\mathbb{Z}, \ldots, (m-1)+m\mathbb{Z}$.

Next we consider the following important example. The vector space $\mathbb{R}^3$ is clearly an abelian group under addition. Let L be an arbitrary plane that passes through the origin. Such a plane can be specified by an equation of the

form $Ax+By+Cz=0$ for certain real numbers A,B,C and we will suppose, for convenience, that $C\neq 0$. We note that L is a subgroup of the additive group $\mathbb{R}^3$. This can be shown as follows. Certainly $L\neq\emptyset$. Let $(x_1,y_1,z_1),(x_2,y_2,z_2)\in L$. Then $Ax_1+By_1+Cz_1=0=Ax_2+By_2+Cz_2$ and we have

$$\begin{aligned}&A(x_1-x_2)+B(y_1-y_2)+C(z_1-z_2)\\&\quad=(Ax_1+By_1+Cz_1)-(Ax_2+By_2+Cz_2)\\&\quad=0-0=0.\end{aligned}$$

Hence $(x_1-y_1,x_2-y_2,x_3-y_3)\in L$, so L satisfies the condition (**ASG 3**) and L is a subgroup by Corollary 3.1.5.

Now let K be an arbitrary plane which is parallel to L. Its points will satisfy an equation of the form $Ax+By+Cz=D$ for some $D\in\mathbb{R}$. Thus, if $(x_3,y_3,z_3)\in K$, then $Ax_3+By_3+Cz_3=D$, and therefore, since $C\neq 0$, we have

$$\begin{aligned}(x_3,y_3,z_3)&=(x_3,y_3,(D-Ax_3-By_3)/C)\\&=(x_3,y_3,-(Ax_3+By_3)/C)+(0,0,D/C).\end{aligned}$$

We note that $Ax_3+By_3+C(-(Ax_3+By_3)/C)=0$. Thus every element of K has the form $\alpha+\delta$, where $\alpha\in L$, $\delta=(0,0,D/C)$, so K is an affine subspace defined by the subspace L.

These two examples are very similar in nature and lead us to the following general definition.

Definition 3.2.1. *Let G be a group and let H be a subgroup of G. For each $x\in G$ the subset $xH=\{xu\mid u\in H\}$ is called a left coset of H in G, or a left H-coset, and the element x is called a representative of this coset.*

Similarly $Hx=\{ux\mid u\in H\}$ is called a right coset of H in G, or a right H-coset, and the element x is called a representative of this coset.

In additive notation xH is written $x+H=\{x+u|u\in H\}$ and the corresponding right coset will be $H+x$. This shows that in the previous two examples we are actually constructing cosets of certain subgroups. Of course in the case of abelian groups $xH=Hx$ since $hx=xh$. Thus, for abelian groups, the right and left cosets are identical.

However it is not always the case that $xH=Hx$, as can be seen from the following simple example.

Example. Consider the group $\mathbf{S}_3$ and its subgroup $T_1=\{\iota_{12},\varepsilon\}$, as seen in Section 3.1. The Cayley table for the group $\mathbf{S}_3$, can be used to show:

$$\varepsilon T_1=T_1=T_1\varepsilon,$$

$$\iota_{13}T_1 = \{\iota_{13}\iota_{12}, \iota_{13}\varepsilon\} = \{\lambda, \iota_{13}\} = \lambda T_1,$$

$$T_1\iota_{13} = \{\iota_{12}\iota_{13}, \varepsilon\iota_{13}\} = \{\mu, \iota_{13}\} = T_1\mu,$$

and

$$\iota_{23}T_1 = \{\mu, \iota_{23}\} = \mu T_1, T_1\iota_{23} = \{\lambda, \iota_{23}\} = T_1\lambda.$$

These calculations clearly show that it is not always the case that $xH = Hx$. In the sequel we shall usually state and prove results for left cosets, but there will usually be an analogous result for right cosets, which we invite the reader to verify. Our next few results could be phrased in the language of equivalence relations, but here we give direct proofs.

Proposition 3.2.2. *Let G be a group, let H be a subgroup of G and let $x \in G$. If $y \in xH$, then $xH = yH$.*

Proof. Since $y \in xH$, there is an element $u_1 \in H$ such that $y = xu_1$. If $z \in yH$, then $z = yu_2$ for some $u_2 \in H$, and we have $z = yu_2 = (xu_1)u_2 = x(u_1u_2)$. Since H is a subgroup, the condition (**SG 1**) implies that $u_1u_2 \in H$, so that $z \in xH$. It follows that $yH \subseteq xH$.

Conversely, let w be an arbitrary element of xH. Then $w = xu_3$ for some $u_3 \in H$. We have

$$x = xe = x(u_1u_1^{-1}) = (xu_1)u_1^{-1} = yu_1^{-1}.$$

It follows that $w = xu_3 = (yu_1^{-1})u_3 = y(u_1^{-1}u_3)$. Since H is a subgroup, the condition (**SG 3**) implies that $u_1^{-1}u_3 \in H$, so that $w \in yH$. It follows that $xH \subseteq yH$. Together with the inclusion $yH \subseteq xH$, this means that $xH = yH$.

Proposition 3.2.2 therefore shows that left cosets either are equal or disjoint (and likewise for right cosets).

Corollary 3.2.3. *Let G be a group and let H be a subgroup of G. Then the family of all left H-cosets is a partition of G.*

Proof. If g is an arbitrary element of G, then $g = ge \in gH$. Hence $G = \bigcup_{g \in G} gH$. Suppose that $xH \cap yH \neq \emptyset$ and let $z \in xH \cap yH$. Then $z \in xH$ and Proposition 3.2.2 implies that $zH = xH$. Similarly, $z \in yH$ and we deduce that $zH = yH$. It follows that $xH = yH$. This shows that the distinct left cosets partition G.

If $H = \langle e \rangle$, then $xH = \{x\}$ for each element $x \in G$, so we obtain the largest partition of G, consisting of one-element sets. If $H = G$, then we obtain the smallest partition having only one subset, exactly the group G.

As we saw in Section 1.5, each partition corresponds to an equivalence relation. Now we will discover the equivalence relations that correspond to these partitions.

If G is a group and H is a subgroup of G, then we define a relation Σ_H on G as follows:

$$\text{For all } x, y \in G, \text{ then } (x, y) \in \Sigma_H \text{ if and only if } y^{-1}x \in H.$$

The relation Σ_H is reflexive since if $x \in G$ then $x^{-1}x = e \in H$, so $(x, x) \in \Sigma_H$. The relation Σ_H is symmetric. To see this note that if $x, y \in G$ and $(x, y) \in \Sigma_H$ then $y^{-1}x \in H$. Since H is a subgroup, it contains $(y^{-1}x)^{-1} = x^{-1}(y^{-1})^{-1} = x^{-1}y$ so $(y, x) \in \Sigma_H$ also. The relation Σ_H is also transitive. We let $(x, y), (y, z) \in \Sigma_H$. Then, by definition, $y^{-1}x, z^{-1}y \in H$. Since H is a subgroup, it contains the product $(z^{-1}y)(y^{-1}x) = z^{-1}x$, and hence $(x, z) \in \Sigma_H$.

Consequently, Σ_H is an equivalence relation and we now find the equivalence classes corresponding to this relation. If $(x, y) \in \Sigma_H$, then $y^{-1}x = h \in H$. Multiplying both sides of this equation first by y on the left, and then by h^{-1} on the right we see that $y = xh^{-1} \in xH$. Thus each element which is equivalent to x (relative to the relation Σ_H) belongs to xH. Conversely, if $z \in xH$, then $z = xh$ for some element $h \in H$. We now have

$$z^{-1}x = (xh)^{-1}x = h^{-1}x^{-1}x = h^{-1} \in H,$$

and therefore $(x, z) \in \Sigma_H$. So the equivalence classes of Σ_H are exactly the left cosets of H.

Similarly, there is a relation Λ_H, defined by the rule: If $x, y \in G$, then $(x, y) \in \Lambda_H$ if and only if $xy^{-1} \in H$. As above it can be shown that the equivalence classes under the relation Λ_H are exactly the right cosets of H.

There is a very easy connection between the left and right cosets.

Proposition 3.2.4. *Let G be a group and let H be a subgroup of G. The mapping*

$$\nu : Hx \longmapsto x^{-1}H$$

is a bijection from the set of all right cosets of H onto the set of all left cosets of H.

Proof. First we must show that ν is a mapping, in the sense that it does not depend on the choice of the representative x. To this end, let y be another representative of the coset Hx. Then $Hx = Hy$ so $y = hx$ for some element $h \in H$. Then $y^{-1} = x^{-1}h^{-1} \in x^{-1}H$, and so $y^{-1}H = x^{-1}H$, since left cosets are equal or disjoint. Hence ν is well-defined.

Furthermore, the mapping ν is injective. For, if Hx, Hy are right cosets and if $\nu(Hx) = \nu(Hy)$ then

$$x^{-1}H = \nu(Hx) = \nu(Hy) = y^{-1}H.$$

Then $y^{-1} = x^{-1}h$ for some element $h \in H$, and $y = h^{-1}x \in Hx$. It follows that $Hx = Hy$ and that ν is injective follows.

Finally, ν is surjective since if zH is a left coset then

$$\nu(Hz^{-1}) = (z^{-1})^{-1}H = zH.$$

Hence ν is a bijective mapping, as required.

If G is a group and H is a subgroup of G we choose a representative from each left coset of H in G and denote the collection of these representatives by $\mathbf{lt}(G,H)$. Likewise we choose a representative from each right coset of H in G and denote the resulting set of representatives by $\mathbf{rt}(G,H)$.

Definition 3.2.5. *Let G be a group and H be a subgroup of G.The set $\mathbf{lt}(G,H)$ (respectively $\mathbf{rt}(G,H)$) is called a left transversal or a set of left coset representatives (respectively right transversal or a set of right coset representatives) to H in G.*

Consequently, $G = \bigcup_{x \in \mathbf{lt}(G,H)} xH$ (respectively, $G = \bigcup_{x \in \mathbf{rt}(G,H)} Hx$). Furthermore, the equation $xH = yH$ (respectively $Hx = Hy$) for $x, y \in \mathbf{lt}(G,H)$ (respectively, $x, y \in \mathbf{rt}(G,H)$) means that $x = y$.

Definition 3.2.6. *Let G be a group and let H be a subgroup of G. If the number of distinct right cosets of H in G is finite then H is said to have finite index in G. The number of distinct right cosets of H in G is called the index of H in G and it is denoted by $|G : H|$.*

Proposition 3.2.4 implies that the index is also the number of distinct left cosets of H in G. Thus the index of H can be determined by counting the number of distinct left cosets of H in G.

In other words,

$$|G : H| = |\mathbf{rt}(G,H)| = |\mathbf{lt}(G,H)|.$$

In the case when $H = G$, $|G : H| = 1$. If G is a finite group and $H = \langle e \rangle$, then $|G : H| = |G|$.

Definition 3.2.7. *Let G be a group and let H be a subgroup of G. We say that H has infinite index in G if the set of all right cosets of H in G is infinite.*

If G is a finite group, then every subgroup of G has finite index. If G is an infinite group then it has at least one subgroup, having infinite index, namely $\langle e \rangle$. There is an infinite group G in which every nontrivial subgroup has finite index, namely the group $\mathbb{Z}$. Indeed, by Theorem 3.1.12, if H is a nonzero subgroup of $\mathbb{Z}$, then $H = m\mathbb{Z}$ for some positive integer m. As we saw above, $\mathbb{Z} = \bigcup_{0 \le r < m} (r + m\mathbb{Z})$, which shows that $|\mathbb{Z} : m\mathbb{Z}| = m$ is finite.

We now consider some further examples.

1. Earlier we considered the partition of $\mathbf{S}_3$ into cosets using a subgroup generated by a transposition. Another interesting example is also associated with the permutation group $\mathbf{S}_n$. We point out the following natural decomposition of this group. Let

$$P_i = \{\pi \in \mathbf{S}_n \mid \pi(n) = i\}.$$

Since the elements of $\mathbf{S}_n$ are permutations it follows that $P_i \cap P_j = \emptyset$ whenever $i \neq j$ and that

$$\mathbf{S}_n = \bigcup_{1 \le i \le n} P_i = P_1 \cup P_2 \cup \cdots \cup P_n.$$

The subsets that make up this disjoint union are the right cosets of the subgroup P_n in $\mathbf{S}_n$. First we show that P_n is a subgroup and to this end, let $\lambda, \pi \in P_n$. Then

$$(\lambda \circ \pi)(n) = \pi(\lambda(n)) = \pi(n) = n,$$

which shows that P_n satisfies the condition (**SG 1**). Also, since $\pi \in P_n$, we have

$$n = \varepsilon(n) = (\pi \circ \pi^{-1})(n) = \pi^{-1}(\pi(n)) = \pi^{-1}(n),$$

and hence P_n satisfies the condition (**SG 2**). This shows that P_n is a subgroup of $\mathbf{S}_n$.

Furthermore, if $\sigma \in P_j$, then we have

$$(\sigma \circ \iota_{jn})(n) = \iota_{jn}(\sigma(n)) = \iota_{jn}(j) = n,$$

so that $\sigma \circ \iota_{jn} \in P_n$. It follows that

$$\sigma = \sigma \circ \varepsilon = \sigma \circ (\iota_{jn} \circ \iota_{jn}) = (\sigma \circ \iota_{jn}) \circ \iota_{jn} \in P_n \circ \iota_{jn}.$$

This shows that $P_j \subseteq P_n \circ \iota_{jn}$. Conversely, if $\pi \in P_n$, then

$$(\pi \circ \iota_{jn})(n) = \iota_{jn}(\pi(n)) = \iota_{jn}(n) = j,$$

and hence $P_n \circ \iota_{jn} \subseteq P_j$. Thus $P_n \circ \iota_{jn} = P_j$.

2. We next consider the additive group of real numbers. If $\alpha \in \mathbb{R}$, then there exists an integer k such that $k \leq \alpha \leq k+1$. It follows that $0 \leq \alpha - k \leq 1$. Hence $\alpha \in \beta + \mathbb{Z}$ for some $\beta \in [0,1)$. This shows that the subset $[0,1)$ is a transversal to $\mathbb{Z}$ in $\mathbb{R}$.
3. In Section 3.1 we noted that the set $\mathbb{Q}\backslash\{0\} = \mathbb{Q}^*$ of all nonzero rationals is an abelian group under multiplication and the subset H of all positive rationals is a subgroup of $\mathbb{Q}^*$. If $x \in \mathbb{Q}^*$ and $x < 0$, then $x = (-1)(-x)$. Since $-x > 0$ we have $-x \in H$ which shows that the subset $\{1, -1\}$ is a transversal to H in $\mathbb{Q}^*$.
4. In Section 3.1 we saw that the set $\mathbb{C}\backslash\{0\} = \mathbb{C}^*$ of all nonzero complex numbers is an abelian group under multiplication and the subset $\mathbb{T}_1 = \{\alpha \in \mathbb{C} \mid |\alpha| = 1\}$ is a subgroup. If α is an arbitrary complex number, then $\alpha = r(\cos\varphi + i\sin\varphi)$ where $\cos\varphi + i\sin\varphi \in \mathbb{T}_1$ and r is a positive real number. This shows that the subset of all positive real numbers is a transversal to the subgroup $\mathbb{T}_1$ in $\mathbb{C}^*$.
5. In Section 3.1 it was shown that the subset $\mathbf{GL}_n(\mathbb{R})$ of all nonsingular matrices in $\mathbf{M}_n(\mathbb{R})$ is a group under matrix multiplication, and that the subset $\mathbf{SL}_n(\mathbb{R}) = \{A \in \mathbf{M}_n(\mathbb{R}) \mid \mathbf{det}(A) = 1\}$ is a subgroup of $\mathbf{GL}_n(\mathbb{R})$. For every nonzero real number α define the matrix $\mathbf{d}(\alpha) \in \mathbf{GL}_n(\mathbb{R})$ by

$$\begin{pmatrix} \alpha & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Clearly $\mathbf{det}(\mathbf{d}(\alpha)) = \alpha$. Let $\alpha \in \mathbb{R}^*$ and let A be a nonsingular matrix such that $\mathbf{det}(A) = \alpha$. Let $A_1 = \mathbf{d}(\alpha^{-1})A$. Then

$$\mathbf{det}(A_1) = \mathbf{det}(\mathbf{d}(\alpha^{-1})A) = \mathbf{det}(\mathbf{d}(\alpha^{-1}))\mathbf{det}(A) = \alpha^{-1}\alpha = 1,$$

so that $A_1 \in \mathbf{SL}_n(\mathbb{R})$. However, $A = \mathbf{d}(\alpha)\mathbf{d}(\alpha^{-1})A = \mathbf{d}(\alpha)A_1$. This shows that the subset $\{\mathbf{d}(\alpha) \mid \alpha \in \mathbb{R}\backslash\{0\}\}$ is a transversal to $\mathbf{SL}_n(\mathbb{R})$ in $\mathbf{GL}_n(\mathbb{R})$.

For finite groups the following theorem plays an extremely important role.

Theorem 3.2.8. *(Lagrange's Theorem) Let G be a finite group and let H be a subgroup of G. Then $|G| = |G : H| \cdot |H|$. In particular, the order of a subgroup of a finite group is a divisor of the order of the group.*

Proof. Let $L = \mathbf{lt}(G, H)$ be a left transversal to H in G. Then $|L| = |G : H|$. Furthermore, the family $\{xH \mid x \in L\}$ is a partition of G, and therefore $|G| = \sum_{x \in L} |xH|$.

Let $x \in G$ and consider the mapping $f : H \longrightarrow xH$, defined by $f(h) = xh$, for each $h \in H$. This mapping is surjective, by definition. It is also injective as we now show. If $h_1, h_2 \in H$ and $f(h_1) = f(h_2)$ then $xh_1 = xh_2$. We multiply both sides of this equation on the left by x^{-1} and note that $x^{-1}x = e$. This gives $x^{-1}(xh_1) = x^{-1}(xh_2)$ so $eh_1 = eh_2$ and hence $h_1 = h_2$. Since f is surjective and injective, it follows that f is bijective. Hence $|H| = |xH|$, for each $x \in G$.

Now we have

$$|G| = \sum_{x \in L} |xH| = |L|\,|H| = |G : H|\,|H|,$$

since the cosets xH, for $x \in L$, are disjoint. This proves Lagrange's theorem.

We make the further obvious remark that if G is a finite group and $H \leq G$ then $|G : H| = |G|/|H|$. A further immediate consequence is as follows.

Corollary 3.2.9. *Let G be a finite group and let x be an element of G. Then the order of x is a divisor of the order of G.*

To see this we recall from Section 3.1 that $|x| = |\langle x \rangle|$.

Corollary 3.2.10. *Let G be a finite group. If $|G|$ is a prime, then G is a cyclic group.*

Proof. Let $e \neq g \in G$. Then $\langle g \rangle$ has at least one nontrivial element, so that $|\langle g \rangle| > 1$. Since $|G|$ is a prime, Lagrange's Theorem implies that $|\langle g \rangle| = |G|$, and hence $\langle g \rangle = G$. This proves the result.

We have already seen (in the group $\mathbf{S}_3$) that there are subgroups, whose left cosets do not coincide with the right cosets. However, there are subgroups for which the left cosets and the right cosets coincide. For example, this will always be the case in an abelian group and it happens in other cases also. It is natural to consider such subgroups in more detail. We point out at once one other type of subgroup where this property is easily verified.

Proposition 3.2.11. *Let G be a group and let H be a subgroup of G. If $|G : H| = 2$, then $xH = Hx = G \setminus H$ for each element $x \notin H$ and $xH = Hx = H$ for each element $x \in H$.*

Proof. Indeed, we have $G = H \cup gH$ for some element $g \in G$, since $|G : H| = 2$. It follows that $gH = G \setminus H$. If $x \notin H$ then $xH \neq H$ and hence $xH = gH = G \setminus H$. A similar argument is valid for right cosets. Thus $xH = Hx = G \setminus H$ when $x \notin H$. On the other hand if $x \in H$ then $xH = H = Hx$ and the result follows.

Now we turn back to the group $\mathbf{S}_3$ and consider its subgroup $L = \{\lambda, \mu, \varepsilon\}$ (see Section 3.1). Since $|L| = 3$, Lagrange's theorem implies that $|\mathbf{S}_3 : L| = 2$, and using Proposition 3.2.11 we deduce that the left cosets and right cosets of L coincide.

We make the following remark. Suppose that G is a group and H is a subgroup of G such that the family of all left cosets of H in G coincides with the family of all right cosets of H in G. This means that if $x \in G$, then there exists $y \in G$ such that $xH = Hy$. Since $x \in xH$, we have $x \in Hy$. Since $x \in Hx \cap Hy$ and since right cosets are equal or disjoint, we have that $Hy = Hx$ so $xH = Hx$. Thus, in this case, each individual left coset xH coincides with is corresponding right coset Hx and this helps prompt the following definition.

Definition 3.2.12. *Let G be a group. The subgroup H is called normal in G, if $xH = Hx$ for each element $x \in G$. We denote the fact that H is normal in G by $H \lhd G$.*

Note that every group G automatically has at least two normal subgroups since both the trivial subgroup $\langle e \rangle$ and the entire group G are normal in G. We note also that normality is always relative to some larger group. If $H \leq K \leq G$ then it is perfectly possible for H to be normal in K, but not normal in the larger group G. Clearly, however, if H is normal in G then H is also normal in K.

Definition 3.2.13. *A group G is called simple if it has only two normal subgroups, namely $\langle e \rangle$ and G.*

Every group of prime order p is simple since, by Lagrange's Theorem, such a group even has no proper nontrivial subgroups. Furthermore, if G is an abelian simple group and $e \neq x \in G$, then $\langle x \rangle \lhd G$ so $\langle x \rangle = G$. Also if $|G| = mn$, where $m, n \neq 1$, then $\langle e \rangle \neq \langle x^n \rangle \lneq \langle x \rangle$. Hence G must be cyclic of prime order.

Galois showed that the group $\mathbf{A}_5$ is simple and it was later proved that the group $\mathbf{A}_n$ is simple for each $n \geq 5$. This was the first series of finite nonabelian simple groups, and it is now known that there are a number of other such series.

Let G be a group, and let $\zeta(G) = \{x \in G \mid xg = gx \text{ for each } g \in G\}$. The subset $\zeta(G)$ is a subgroup of G. To see this, note that certainly $e \in \zeta(G)$ so $\zeta(G) \neq \emptyset$. Also if $x, y \in \zeta(G)$ and g is an arbitrary element of G, then $(xy)g = x(yg) = x(gy) = (xg)y = (gx)y = g(xy)$, so that $\zeta(G)$ satisfies the condition (**SG 1**). Furthermore, multiplying both sides of the equation $xg = gx$ first on the right and then on the left by x^{-1}, we deduce that $gx^{-1} = x^{-1}g$, which shows that $x^{-1} \in \zeta(G)$. Hence $\zeta(G)$ satisfies the condition (**SG 2**), and this shows that $\zeta(G)$ is a subgroup of G. The subgroup $\zeta(G)$ is called *the center of the group* G. Next we note that every subgroup H of $\zeta(G)$ is normal in G. For if $x \in G$ and $h \in H$, then xH and Hx must coincide because $xh = hx$ whenever $h \in H$ in this case. Since a group G is abelian if and only if $G = \zeta(G)$ this affords a proof that every subgroup of an abelian group is normal.

We note in passing that in general if H is a normal subgroup of a group G then xH and Hx are equal setwise. This does not generally mean that $xh = hx$ for all elements $h \in H$.

It should be noted that the center of a group can be trivial, meaning it consists of the identity only. This is the case for the group $\mathbf{S}_3$, for example.

Definition 3.2.14. *Let G be a group and let $g, x \in G$. We say that the elements g and $x^{-1}gx$ are conjugate in G by the element x.*

We remark that the relation "to be conjugate" is an equivalence relation on G. The relation is reflexive since $g = g^{-1}gg$. It is symmetric because if $v = x^{-1}gx$, then $g = xvx^{-1} = (x^{-1})^{-1}vx^{-1}$. Finally it is transitive because if $v = x^{-1}gx$ and $u = y^{-1}vy$, then

$$u = y^{-1}vy = y^{-1}(x^{-1}gx)y = y^{-1}x^{-1}gxy = (xy)^{-1}g(xy).$$

We write $g^G = \{x^{-1}gx \mid x \in G\}$. Then the subset g^G is an equivalence class under the relation "to be conjugate"and it is called the *conjugacy class of* g. In general the subset g^G is not a group.

Next, for the group G, its subgroup H and an element $x \in G$, let $x^{-1}Hx = \{x^{-1}hx \mid h \in H\}$. We note that the nonempty subset $x^{-1}Hx$ is a subgroup of G. To see this let $u, v \in x^{-1}Hx$. Then $u = x^{-1}h_1x, v = x^{-1}h_2x$ for certain elements $h_1, h_2 \in H$ and we have, using Proposition 1.5.7,

$$uv^{-1} = (x^{-1}h_1x)(x^{-1}h_2x)^{-1} = x^{-1}h_1xx^{-1}h_2^{-1}x = x^{-1}h_1h_2^{-1}x.$$

Since H is a subgroup, $h_1h_2^{-1} \in H$, so that $uv^{-1} \in x^{-1}Hx$. Corollary 3.1.5 shows that $x^{-1}Hx$ is a subgroup of G. We say that the subgroups H and $x^{-1}Hx$ are *conjugate in G by the element x*.

Using just these concepts, we can now deduce the following criterion for normal subgroups.

Theorem 3.2.15. *Let G be a group and let H be a subgroup of G. The following are equivalent:*

(i) *H is a normal subgroup of G;*
(ii) *$h^G \subseteq H$ for every element $h \in H$;*
(iii) *$x^{-1}Hx = H$ for every element $x \in G$.*

Proof. (i) $\Longrightarrow$ (ii). Suppose that (i) holds, let h be an arbitrary element of H and let g be an arbitrary element of G. Since H is normal in G, $gH = Hg$. It follows that $hg \in gH$, so that $hg = gh_1$ for some element $h_1 \in H$. Then $g^{-1}hg = h_1 \in H$. Since this is true for each $g \in G$ we have $h^G \subseteq H$.

(ii) $\Longrightarrow$ (iii). Suppose that (ii) holds. Let $u \in x^{-1}Hx$, so $u = x^{-1}hx$ for some $h \in H$. By (ii), $u \in H$, so $x^{-1}Hx \subseteq H$. Conversely, suppose that $y \in H$. We have $y = (x^{-1}x)y(x^{-1}x) = x^{-1}(xyx^{-1})x$. Clearly, $xyx^{-1} = (x^{-1})^{-1}yx^{-1} \in y^G$, and using (ii) we deduce that $xyx^{-1} = y_1 \in H$. Then $y = x^{-1}y_1x \in x^{-1}Hx$, so $H \subseteq x^{-1}Hx$ and (iii) follows.

(iii) $\Longrightarrow$ (i). Finally suppose that (iii) holds. We have to show that $xH = Hx$ for all $x \in G$. To this end, let $z \in xH$. Then $z = xh_2$ for some element $h_2 \in H$. By (iii), there is an element $h_3 \in H$ such that $h_2 = x^{-1}h_3x$. It follows that $z = xh_2 = x(x^{-1}h_3x) = (xx^{-1})h_3x = h_3x \in Hx$ and we deduce that $xH \subseteq Hx$. Suppose now that $w \in Hx$. Then $w = h_4x$ for some element $h_4 \in H$. Using (iii) again, we see that there exists an element $h_5 \in H$ such that $h_4 = xh_5x^{-1} = (x^{-1})^{-1}h_5x^{-1}$. It follows that $w = h_4x = xh_5x^{-1}x = xh_5 \in xH$. Hence $Hx \subseteq xH$ and therefore $Hx = xH$. The result now follows.

We note the following important subgroup property.

Proposition 3.2.16. *Let G be a group and let $\mathfrak{S}$ be a family of normal subgroups of G. Then the intersection, $\cap\mathfrak{S}$, of all subgroups of this family is also normal in G.*

Proof. Let $S = \cap\mathfrak{S}$. By Proposition 3.1.8, S is a subgroup of G. Let $x \in S$ and $g \in G$. If U is an arbitrary subgroup of the family $\mathfrak{S}$ then $x \in U$ and, since U is normal in G, U also contains the element $g^{-1}xg$. It follows that $g^{-1}xg$ belongs to the intersection of all subgroups of the family $\mathfrak{S}$ and hence to S. Therefore $x^G \subseteq S$ and Theorem 3.2.15 completes the proof.

We now give some examples of normal subgroups. As we have already mentioned, every subgroup H of an abelian group G is a normal subgroup of G.

By Proposition 3.2.11 every subgroup of a group G having index 2 is normal in G. In particular, it follows that the subgroup $\mathbf{A}_n$ of all even permutations is always a normal subgroup of $\mathbf{S}_n$.

Next we consider the group $\mathbf{GL}_n(\mathbb{R})$ of all nonsingular $n \times n$ matrices. In Section 3.1 we proved that $\mathbf{SL}_n(\mathbb{R}) = \{A \in \mathbf{M}_n(\mathbb{R}) \mid \mathbf{det}(A) = 1\}$ is a subgroup of $\mathbf{GL}_n(\mathbb{R})$. Now we show that this subgroup is normal. Indeed, let $A \in \mathbf{SL}_n(\mathbb{R})$ and $B \in \mathbf{GL}_n(\mathbb{R})$. Then

$$\begin{aligned}\mathbf{det}(B^{-1}AB) &= \mathbf{det}(B^{-1})\mathbf{det}(A)\mathbf{det}(B) = \frac{1}{\mathbf{det}(B)}\mathbf{det}(B)\mathbf{det}(A)\\ &= \mathbf{det}(A) = 1,\end{aligned}$$

which implies that $B^{-1}AB \in \mathbf{SL}_n(\mathbb{R})$. Theorem 3.2.15 shows that $\mathbf{SL}_n(\mathbb{R})$ is a normal subgroup of $\mathbf{GL}_n(\mathbb{R})$.

In Section 3.1 we noted that the subset $\mathbf{T}_n^0(\mathbb{R})$ of all nonsingular upper-triangular matrices is a subgroup of $\mathbf{GL}_n(\mathbb{R})$. Let $A = (a_{jm}), B = (b_{jm}) \in \mathbf{T}_n^0(\mathbb{R})$, and let $AB = (c_{jm})$ where $1 \leq j, m \leq n$. For the diagonal coefficient c_{jj} we have

$$c_{jj} = a_{j1}b_{1j} + \cdots + a_{j,j-1}b_{j-1,j} + a_{jj}b_{jj} + a_{j,j+1}b_{j+1,j} + \cdots + a_{jn}b_{nj}.$$

Since $a_{j1} = \cdots = a_{j,j-1} = b_{j+1,j} = \cdots = b_{nj} = 0$, $c_{jj} = a_{jj}b_{jj}$, for $1 \leq j \leq n$. In particular, if $B = A^{-1}$, then $c_{jj} = 1$ and $b_{jj} = \frac{1}{a_{jj}}$, for $1 \leq j \leq n$. The triangular matrix $A = (a_{jm})$ is called *unitriangular* if $a_{jj} = 1$ for all j, where $1 \leq j \leq n$. Let $\mathbf{UT}_n(\mathbb{R})$ denote the set of all unitriangular matrices in $\mathbf{T}_n^0(\mathbb{R})$. We show that the subset $\mathbf{UT}_n(\mathbb{R})$ is a normal subgroup of $\mathbf{T}_n^0(\mathbb{R})$. To this end let $A = (a_{jm}), B = (b_{jm}) \in \mathbf{UT}_n(\mathbb{R})$ and let $AB = (c_{jm})$. By what we saw above, $c_{jj} = a_{jj}b_{jj} = 1 \cdot 1 = 1$, for $1 \leq j \leq n$, so that $AB \in \mathbf{UT}_n(\mathbb{R})$. Since A is a triangular matrix, $A^{-1} = D = (d_{jm}) \in \mathbf{T}_n^0(\mathbb{R})$ and, as above, we obtain $d_{jj} = \frac{1}{a_{jj}} = 1$, for $1 \leq j \leq n$. Thus $A^{-1} \in \mathbf{UT}_n(\mathbb{R})$ and $\mathbf{UT}_n(\mathbb{R})$ satisfies conditions (**SG 1**) and (**SG 2**). Hence $\mathbf{UT}_n(\mathbb{R})$ is a subgroup of $\mathbf{T}_n^0(\mathbb{R})$. Next we show that $\mathbf{UT}_n(\mathbb{R})$ is a normal subgroup of $\mathbf{T}_n^0(\mathbb{R})$. Let $U = (u_{jm}) \in \mathbf{UT}_n(\mathbb{R})$, $A = (a_{jm}) \in \mathbf{T}_n^0(\mathbb{R})$, and $A^{-1} = (d_{jm})$. We note that $u_{jj} = 1$, for $1 \leq j \leq n$. Since $\mathbf{T}_n^0(\mathbb{R})$ is a subgroup, $A^{-1}UA = (v_{jm}) \in \mathbf{T}_n^0(\mathbb{R})$. Again by the work above, $v_{jj} = (\frac{1}{a_{jj}})u_{jj}a_{jj} = (\frac{1}{a_{jj}})a_{jj} = 1$, for $1 \leq j \leq n$. Hence $A^{-1}UA \in \mathbf{UT}_n(\mathbb{R})$. Using Theorem 3.2.15, we deduce that $\mathbf{UT}_n(\mathbb{R})$ is a normal subgroup of $\mathbf{T}_n^0(\mathbb{R})$.

Finally in this section we consider a very specific example. In the group $\mathbf{GL_4}(\mathbb{R})$ consider the matrices

$$e = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad a = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$b = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \quad c = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}.$$

Let $Q = \{e, -e, a, -a, b, -b, c, -c\}$. The following multiplication table can be easily obtained using matrix multiplication.

	e	$-e$	a	$-a$	b	$-b$	c	$-c$
e	e	$-e$	a	$-a$	b	$-b$	c	$-c$
$-e$	$-e$	e	$-a$	a	$-b$	b	$-c$	c
a	a	$-a$	$-e$	e	c	$-c$	$-b$	b
$-a$	$-a$	a	e	$-e$	$-c$	c	b	$-b$
b	b	$-b$	$-c$	c	e	$-e$	a	$-a$
$-b$	$-b$	b	c	$-c$	$-e$	e	$-a$	a
c	c	$-c$	b	$-b$	$-a$	a	$-e$	e
$-c$	$-c$	c	$-b$	b	a	$-a$	e	$-e$

Using this table, we can see that Q satisfies the conditions (**SG 1**), (**SG 2**) and hence it is a subgroup of $\mathbf{GL_4}(\mathbb{R})$. This group is called the *Quaternion group*. It is easy to see that the elements $a, -a, b, -b, c, -c$ each have order 4, and that $\langle a \rangle = \langle -a \rangle, \langle b \rangle = \langle -b \rangle, \langle c \rangle = \langle -c \rangle$. Every cyclic subgroup of order 4 has index 2 in Q so Proposition 3.2.11 implies that $\langle a \rangle, \langle b \rangle$, and $\langle c \rangle$ are each normal in Q. The element $-e$ has order 2 and $\langle -e \rangle = \langle a \rangle \cap \langle b \rangle = \langle a \rangle \cap \langle c \rangle = \langle b \rangle \cap \langle c \rangle$. It follows that $-e$ commutes with a, b, c so $-e \in \zeta(Q)$. Hence $\langle -e \rangle$ is also normal in Q. A little consideration shows that Q has no other proper, nontrivial subgroups. Thus every subgroup of Q is normal. At the same time, Q is nonabelian. For example, $ab = c$, but $ba = -c \neq c$.

A group G is said to be a *Dedekind group* if every subgroup of G is normal. These groups were so named after the German mathematician R. Dedekind who, in 1897, obtained the description of those finite groups all of whose subgroups are normal. Dedekind groups are, in a sense, the exact opposite of simple groups. Of course, all abelian groups are Dedekind and Dedekind showed that Dedekind groups are not too far removed from the examples we have considered. The Quaternion group gives us an example of a nonabelian Dedekind group.

Exercise Set 3.2

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

3.2.1. Let $G = \langle g \rangle, |g| = 16$. Find all elements $x \in G$ with the property $G = \langle x \rangle$.

3.2.2. Let $G = \langle g \rangle, |g| = 15$. Find all elements $x \in G$ with the property $G = \langle x \rangle$.

3.2.3. Prove that the following set of matrices,

$$N = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \right\},$$

is a subgroup of $\mathbf{GL}_2(\mathbb{Z})$. Is this subgroup normal in $\mathbf{GL}_2(\mathbb{Z})$?

3.2.4. Prove that the following set of matrices,

$$N = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \right\},$$

is a normal subgroup of the subgroup $\mathbf{D}_2(\mathbb{Z})$ of all diagonal matrices in $\mathbf{GL}_2(\mathbb{Z})$.

3.2.5. Let $H = \{\pi \mid \pi \in \mathbf{A}_5, \pi(5) = 5\}$. Prove that H is a subgroup of $\mathbf{A}_5$. Find $|H|$.

3.2.6. Let $H = \{\pi \mid \pi \in \mathbf{A}_5, \pi(5) = 5, \pi(4) = 4\}$. Prove that H is a subgroup of $\mathbf{A}_5$. Find $|H|$.

3.2.7. Find the order of the element $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ in the group $\mathbf{GL}_2(\mathbb{Q})$.

3.2.8. Find the order of the element $A = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$ in the group $\mathbf{GL}_2(\mathbb{Q})$.

3.2.9. Prove that if a group G is a cyclic group of order p^2, where p is a prime, then G has exactly three subgroups.

3.2.10. Let G be a cyclic group of order 12. How many subgroups does this group have?

3.2.11. Prove that $Z(\mathbf{S}_n)$, the center of $\mathbf{S}_n$ is trivial for all $n \geq 3$.

3.2.12. Let n be a natural number and let $(a_1\, a_2 \ldots a_r)$ be a cycle of length r in $\mathbf{S}_n$. Let $\theta \in \mathbf{S}_n$. Prove that $\theta^{-1}(a_1\, a_2 \ldots a_r)\theta = (\theta(a_1)\, \theta(a_2) \ldots \theta(a_r))$. Deduce that if $\alpha \in \mathbf{S}_n$ and if $\alpha = \alpha_1\alpha_2 \ldots \alpha_k$, as a product of disjoint cycles, then $\theta^{-1}\alpha\theta = \gamma_1\gamma_2 \ldots \gamma_k$ has the same cycle decomposition as α.

3.2.13. Let G be a group and suppose that $N = \langle a \rangle$ is a cyclic normal subgroup of G. Prove that every subgroup of N is also normal in G.

3.2.14. Prove that if G is a group, N is a normal subgroup of G and H is a subgroup of G then $H \cap N$ is a normal subgroup of H.

3.2.15. Let n be a natural number. Let $\mathbf{SL}_2(\mathbb{Z})$ denote the subgroup of $\mathbf{GL}_2(\mathbb{R})$ consisting of matrices A with coefficients in $\mathbb{Z}$ and such that $\mathbf{det}\, A = 1$. Let

$$X = \left\{ \begin{pmatrix} 1+na & nb \\ nc & 1+nd \end{pmatrix} \in \mathbf{SL}_2(\mathbb{Z}) | a,b,c,d \in \mathbb{Z} \right\}.$$

Prove that X is a normal subgroup of $\mathbb{SL}_2(\mathbb{Z})$.

3.2.16. Let n be a natural number, at least 3. The group $\mathbf{D}_{2n}$ is the group of symmetries of the regular n-gon. (Thus $\mathbf{D}_8$ is the group of symmetries of a square.) Prove that $\mathbf{D}_{2n}$ has order $2n$, consisting of n rotations and n reflections. Let α be a rotation through $2\pi/n$ radians counterclockwise about the centre of the n-gon and let β be a reflection. Prove that every element of $\mathbf{D}_{2n}$ is either of the form α^j or $\alpha^j\beta$, where $0 \leq j < n$.

3.2.17. Use the notation of the previous problem. Prove that $\langle \alpha \rangle$ is a normal subgroup of $\mathbf{D}_{2n}$ but that $\langle \beta \rangle$ is not normal in $\mathbf{D}_{2n}$.

3.2.18. Show that in $\mathbf{D}_8$ we can find subgroups H, K such that $H \triangleleft K$ and $K \triangleleft \mathbf{D}_8$ but H is not normal in $\mathbf{D}_8$.

3.2.19. Let H be a subgroup of a group G. Let $N = \cap\{g^{-1}Hg | g \in G\}$. Prove that N is a normal subgroup of G.

3.2.20. Let $k \in \mathbb{N}$ be fixed. Let G be a group in which the equation $(xy)^k = x^k y^k$ holds for all $x, y \in G$. Prove that if $H = \{x^k | x \in G\}$ then H is a normal subgroup of G.

3.3 FACTOR GROUPS AND HOMOMORPHISMS

Let G be a group and H a normal subgroup of G. We define a multiplication on the set of all left cosets as follows:

$$xHyH = xyH, \text{ for all } x, y \in G.$$

We have to check that this operation is well-defined. Let $x_1, y_1 \in G$ be elements such that $x_1 H = xH$ and $y_1 H = yH$. Then $x_1 = xh_1$ and $y_1 = yh_2$ for certain elements $h_1, h_2 \in H$. Then $x_1 y_1 = (xh_1)(yh_2) = x(h_1 y)h_2$. Since H is a normal subgroup, $h_1 y \in Hy = yH$, so that $h_1 y = yh_3$ for some $h_3 \in H$, and we have $x_1 y_1 = x(h_1 y)h_2 = x(yh_3)h_2 = (xy)(h_3 h_2) \in xyH$. As we saw in Section 3.2, it follows that $x_1 y_1 H = xyH$, so the multiplication is well-defined.

This operation is associative since, using the associative law in G, we have

$$(xHyH)zH = xyHzH = (xy)zH = x(yz)H = xHyzH = xH(yHzH).$$

The identity element is H itself since

$$(xH)H = xHeH = (xe)H = xH, \text{ and } H(xH) = eHxH = exH = xH.$$

The multiplicative inverse of xH is $x^{-1}H$ because

$$(xH)(x^{-1}H) = xx^{-1}H = eH = H,$$

and

$$(x^{-1}H)(xH) = x^{-1}xH = eH = H.$$

Thus the set $\{xH \mid x \in G\}$ satisfies all the conditions of Definition 3.1.1 and hence it is a group.

Definition 3.3.1. *Let G be a group and let H be a normal subgroup of G. The group of all left H-cosets is called the factor (or quotient) group of G by H and it is denoted by G/H.*

It is interesting to note that some properties of a group are inherited by its factor groups. For example, if G is an abelian group, then each of its factor groups is also abelian. To see this note that if H is normal in G and if $x, y \in G$ then $xHyH = xyH = yxH = yHxH$. It is obvious also that every factor group of a finite group is finite. However some properties of a group are not inherited by factor groups as happens, for example, with the property of being infinite. **For example**, consider the additive group $\mathbb{Z}$ of all integers. If H is a subgroup of $\mathbb{Z}$, then as we saw in Section 3.1, $H = n\mathbb{Z}$ for some $n \geq 0$. In particular, if $H \neq \langle 0 \rangle$ subgroup, then $n > 0$. As we saw in Section 3.2, in this case,

$$\mathbb{Z}/n\mathbb{Z} = \{k + n\mathbb{Z} \mid k \in \mathbb{Z}\} = \{0 + n\mathbb{Z} = n\mathbb{Z}, 1 + n\mathbb{Z}, \ldots, n - 1 + n\mathbb{Z}\}$$

so that $|\mathbb{Z}/n\mathbb{Z}| = n$. Hence $\mathbb{Z}$ is an infinite group, but all its nontrivial subgroups have finite index.

Definition 3.3.2. *The factor group G/H is called proper if H is a nontrivial normal subgroup.*

If $H = \langle e \rangle$, then for each element $x \in G$ we have $xH = x\langle e \rangle = \{x\}$, and $xHyH = \{x\}\{y\} = \{xy\}$. This means that the factor group $G/\langle e \rangle$ is no different from G. In particular the algebraic properties of G and $G/\langle e \rangle$ are identical.

The following important type of mappings are connected to normal subgroups. Let G be a group and H a normal subgroup of G. Define the mapping $\nu_H : G \longrightarrow G/H$ by $\nu_H(x) = xH$ for each element $x \in G$. We have $\nu_H(xy) = xyH = xHyH = \nu_H(x)\nu_H(y)$.

Definition 3.3.3. *Let G, H be groups and let $f : G \longrightarrow H$ be a mapping. Then f is called a homomorphism if*

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

for all $x, y \in G$.

An injective homomorphism is called a monomorphism, a surjective homomorphism is called an epimorphism and a bijective homomorphism is called an isomorphism. If $f : G \longrightarrow H$ is an isomorphism, then Corollary 1.3.5 shows that there is an inverse mapping $f^{-1} : H \longrightarrow G$, which is also bijective. Furthermore f^{-1} is also a homomorphism as we now show. Indeed, let $u, v \in H$ and let $x = f^{-1}(u), y = f^{-1}(v)$. Then $u = f(x), v = f(y)$ and $uv = f(x)f(y) = f(xy)$, since f is a homomorphism, so that $f^{-1}(uv) = xy$. This implies that

$$f^{-1}(u)f^{-1}(v) = xy = f^{-1}(uv),$$

which shows that f^{-1} is a homomorphism and, being bijective, is also an isomorphism.

Definition 3.3.4. *Let G, H be groups. Then G and H are called isomorphic if there exists an isomorphism from G onto H and we write this as $G \cong H$.*

The map $\nu_H : G \longrightarrow G/H$ defined above is a homomorphism called the natural or canonical homomorphism. It is easy to see that the identity mapping $\varepsilon_G : G \longrightarrow G$ is an isomorphism and also that if $f : G \longrightarrow H$ and $g : H \longrightarrow K$ are homomorphisms then their product $g \circ f$ is also a homomorphism.

Proposition 3.3.5. *Let G, U be groups and let $f : G \longrightarrow U$ be a homomorphism. Then*

(i) *$f(e_G) = e_U$ is the identity element of U;*
(ii) *$f(x^{-1}) = f(x)^{-1}$, for all $x \in G$;*
(iii) *If H is a subgroup of G then its image, $f(H)$, is a subgroup of U. In particular, $f(G) = \mathbf{Im} f$ is a subgroup of U;*

(iv) *If V is a subgroup of U then its preimage, $f^{-1}(V)$, is a subgroup of G;*

(v) *If V is a normal subgroup of U then its preimage, $f^{-1}(V)$, is a normal subgroup of G. In particular, $f^{-1}(\langle e_U\rangle)$ is a normal subgroup of G.*

Proof.

(i) We write $e = e_G$. By definition of the identity element, $ex = x$ for every $x \in G$ and hence $ee = e$. It follows that

$$f(e) = f(ee) = f(e)f(e).$$

Since $f(e)$ has an inverse we obtain, multiplying on the right (or left) by $f(e)^{-1}$,

$$e_U = e_U f(e) = f(e).$$

(ii) By definition of inverses, we have $xx^{-1} = e = x^{-1}x$. It follows that

$$f(x)f(x^{-1}) = f(xx^{-1}) = f(e) = e_U = f(e) = f(x^{-1}x) = f(x^{-1})f(x).$$

Thus $f(x)^{-1} = f(x^{-1})$.

(iii) Of course, $f(H) \neq \emptyset$. Let $u, v \in f(H)$. Then there exist elements $a, b \in H$ such that $u = f(a)$ and $v = f(b)$. We have

$$uv^{-1} = f(a)f(b)^{-1} = f(a)f(b^{-1}) = f(ab^{-1}) \in f(H),$$

because H is a subgroup of G. By Corollary 3.1.5, $f(H)$ is a subgroup of U.

(iv) Certainly $f^{-1}(V) \neq \emptyset$. Let $x, y \in f^{-1}(V)$. Then $f(x), f(y) \in V$ and so $f(x)f(y)^{-1} \in V$. Thus

$$f(xy^{-1}) = f(x)f(y^{-1}) = f(x)f(y)^{-1} \in V,$$

and hence $xy^{-1} \in f^{-1}(V)$. By Corollary 3.1.5 again, $f^{-1}(V)$ is a subgroup of G.

(v) Let g be an arbitrary element of G and let $x \in f^{-1}(V)$. Then $f(x) \in V$ and

$$f(g^{-1}xg) = f(g^{-1})f(x)f(g) = (f(g))^{-1}f(x)f(g) \in V,$$

because V is normal in U. It follows that $g^{-1}xg \in f^{-1}(V)$ and by Theorem 3.2.15, $f^{-1}(V)$ is a normal subgroup of G. Finally, we observe that $\langle e_U\rangle \triangleleft U$, so $f^{-1}(\langle e_U\rangle) \triangleleft G$

We saw in (v) that $f^{-1}(\langle e_U\rangle)$ is a normal subgroup of G whenever $f : G \longrightarrow U$ is a homomorphism. This important subgroup is known as the kernel, which we now formally define.

Definition 3.3.6. *Let G, U be groups and let $f : G \longrightarrow U$ be a homomorphism. The normal subgroup $\mathbf{Ker}f = \{x \in G | f(x) = e_U\}$ is called the kernel of the homomorphism f. The subgroup $\mathbf{Im}f = \{f(x) | x \in G\}$ is called the image of f.*

We note that, by contrast with (v), the image of a normal subgroup of G need not be normal in U unless the map f is an epimorphism. As an example, let $j : \mathbf{S}_3 \longrightarrow \mathbf{S}_4$ be the canonical injection, namely the map

$$\begin{pmatrix} 1 & 2 & 3 \\ k_1 & k_2 & k_3 \end{pmatrix} \overset{j}{\longmapsto} \begin{pmatrix} 1 & 2 & 3 & 4 \\ k_1 & k_2 & k_3 & 4 \end{pmatrix}, \text{ where } \{k_1, k_2, k_3\} = \{1, 2, 3\}.$$

Clearly j is a monomorphism. The subgroup $\mathbf{A}_3$ is normal in $\mathbf{S}_3$, but its image is not normal in $\mathbf{S}_4$. To see this notice that

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 3 & 2 \end{pmatrix},$$

which is not an element of $j(\mathbf{A}_3)$.

We next give a number of classical theorems concerning homomorphisms.

Theorem 3.3.7. *(The theorem on monomorphisms). Let G, U be groups. Then a homomorphism $f : G \longrightarrow U$ is a monomorphism if and only if $\mathbf{Ker}f = \langle e\rangle$. In this case, $G \cong \mathbf{Im}f$.*

Proof. Indeed, if f is a monomorphism, then $x \neq e$ implies that $f(x) \neq f(e) = e_U$. This means that no nontrivial element x belongs to $\mathbf{Ker}f$ and hence $\mathbf{Ker}f = \langle e\rangle$.

Conversely, let $\mathbf{Ker}f = \langle e\rangle$, and assume that x, y are elements of G such that $f(x) = f(y)$. Then $f(x)f(y)^{-1} = f(y)f(y)^{-1} = e_U$ and

$$e_U = f(x)f(y)^{-1} = f(x)f(y^{-1}) = f(xy^{-1}),$$

so $xy^{-1} \in \mathbf{Ker}f = \langle e\rangle$. This means that $xy^{-1} = e$ and hence $x = y$. Therefore f is an injective mapping.

For a homomorphism it is often much more convenient to find the kernel in order to determine whether it is a monomorphism or not.

Theorem 3.3.8. *(First Isomorphism Theorem, version 1). Let G, U be groups and let $f : G \longrightarrow U$ be an epimorphism. Then U is isomorphic to the factor group $G/\mathbf{Ker}f$.*

Proof. Let $H = \mathbf{Ker}(f)$ and define a mapping $\Psi_f : G/H \longrightarrow U$, by $\Psi_f(xH) = f(x)$. First we must show that this mapping is well-defined. Let x_1 be an element of G with the property that $x_1H = xH$. Then $x_1 = xh$ for some element $h \in H$ and we have

$$\Psi_f(x_1H) = f(x_1) = f(xh) = f(x)f(h) = f(x)e_U = f(x) = \Psi_f(xH).$$

Thus Ψ_f does not depend on the choice of the representative of the coset xH, so it is well-defined. The mapping Ψ_f is a homomorphism since

$$\Psi_f(xHyH) = \Psi_f(xyH) = f(xy) = f(x)f(y) = \Psi_f(xH)\Psi_f(yH).$$

Furthermore, if $\Psi_f(xH) = e_U$ then the definition of Ψ_f shows that $f(x) = e_U$ and hence $x \in \mathbf{Ker}f = H$. Therefore $xH = H$ so $\mathbf{Ker}f = \langle H \rangle$. By Theorem 3.3.7, Ψ_f is a monomorphism. Finally, let u be an arbitary element of U. Since f is an epimorphism, $u = f(y)$ for some element $y \in G$. Then $u = f(y) = \Psi_f(yH)$, so that Ψ_f is surjective. Thus Ψ_f is an isomorphism.

Theorem 3.3.9. *(First Isomorphism Theorem, version 2) Let G, U be groups and let $f : G \longrightarrow U$ be a homomorphism. Then $G/\mathbf{Ker}f \cong \mathbf{Im}f \leq U$.*

Proof. The restriction of f to the mapping $G \longrightarrow \mathbf{Im}f$ is an epimorphism. Then, by Theorem 3.3.8, we deduce that $\mathbf{Im}f \cong G/\mathbf{Ker}f$. By Proposition 3.3.5, $\mathbf{Im}f$ is a subgroup of U.

We now examine some other interesting examples and as a first application of these theorems we obtain a description of all cyclic groups.

Theorem 3.3.10. *Let $G = \langle g \rangle$ be a cyclic group.*

(i) *If G is infinite then G is isomorphic to the additive group $\mathbb{Z}$ of all integers.*
(ii) *If G is finite and $|G| = m$, then $G \cong \mathbb{Z}/m\mathbb{Z}$.*

Proof. We use multiplicative notation in G, but additive notation in $\mathbb{Z}$. Let $f : \mathbb{Z} \longrightarrow G$ be the mapping defined by $f(n) = g^n$, where $n \in \mathbb{Z}$. We have

$$f(n+k) = g^{n+k} = g^n g^k = f(n)f(k),$$

so that f is a homomorphism. Since every element of G is an integer power of g, f is an epimorphism. Suppose that G is infinite. In this case the integer powers are all distinct since if $g^n = g^k$, with $n > k$ then $g^{n-k} = e$ so $|g|$ is finite, a contradiction. Then $n \neq k$ implies that $f(n) = g^n \neq g^k = f(k)$, so the mapping f is an injection and therefore it is an isomorphism. This proves (i).

Suppose now that G is a finite group. In this case $|g| = m$ and hence $g^m = e$. Thus $m \in \mathbf{Ker} f$ and hence $\langle m \rangle = m\mathbb{Z} \leq \mathbf{Ker} f$. By Theorem 3.1.12, $\mathbf{Ker} f = \langle t \rangle = t\mathbb{Z}$, for some $t \in \mathbb{Z}$, and since $m \in t\mathbb{Z}$, we see that $m = ts$, for some $s \in \mathbb{Z}$. By Theorem 3.3.9, $G \cong \mathbb{Z}/\mathbf{Ker} f$ and hence $|G/\mathbf{Ker} f| = m$. On the other hand, our work from Section 3.2 implies that $|\mathbb{Z}/t\mathbb{Z}| = t$, which shows that $t = m$. Consequently, $G \cong \mathbb{Z}/m\mathbb{Z}$.

Here are some more interesting examples.

1. Let $f_1 : \mathbb{R} \longrightarrow \mathbb{R}^*$ be the mapping defined by $f_1(x) = 13^x$ for each $x \in \mathbb{R}$. We have $f_1(x+y) = 13^{x+y} = 13^x 13^y = f_1(x) f_1(y)$, so that f_1 is a homomorphism. Clearly f_1 is injective and $\mathbf{Im}(f_1) = \{x \mid x \in \mathbb{R} \text{ and } x > 0\} = \mathbb{R}^+$. Thus f_1 is an isomorphism from the group of additive real numbers onto the multiplicative group of all positive real numbers. Of course the inverse map in this case is $\log_{13} : \mathbb{R}^+ \longrightarrow \mathbb{R}$.
2. Define the mapping $f_2 : \mathbb{R}^* \longrightarrow \{1, -1\}$ by

$$f_2(x) = \begin{cases} 1, & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}.$$

   It is easy to check that f_2 is an epimorphism and $\mathbf{Ker}(f_2) = \{x \mid x \in \mathbb{R} \text{ and } x > 0\} = \mathbb{R}^+$. Thus $\mathbb{R}^*/\mathbb{R}^+ \cong \{1, -1\}$ by the First Isomorphism Theorem.
3. Let $f_3 : \mathbb{R}^* \longrightarrow \mathbb{R}^*$ be the mapping defined by $f_3(x) = |x|$ for each $x \in \mathbb{R}^*$. We have $f_3(xy) = |xy| = |x||y| = f_3(x) f_3(y)$, so that f_3 is a homomorphism. It is neither a monomorphism nor an epimorphism since $\mathbf{Ker}(f_3) = \{1, -1\}$ and $\mathbf{Im}(f_3) = \{x | x \in \mathbb{R} \text{ and } x > 0\}$. However the First Isomorphism Theorem shows that $\mathbb{R}^*/\{1, -1\} \cong \mathbb{R}^+$.
4. Next we let $f_4 : \mathbb{R} \longrightarrow \mathbb{C}^*$ denote the mapping, defined by $f_4(\alpha) = \cos\alpha + i \sin\alpha$, for all $\alpha \in \mathbb{R}$. We have, using well-known trigonometric identities or the fact that $f_4(\alpha) = e^{i\alpha}$,

$$\begin{aligned} f_4(\alpha + \beta) &= \cos(\alpha + \beta) + i \sin(\alpha + \beta) = (\cos\alpha + i \sin\alpha)(\cos\beta + i \sin\beta) \\ &= f_4(\alpha) \cdot f_4(\beta), \end{aligned}$$

   so that f_4 is a homomorphism. Again f_4 is neither a monomorphism nor an epimorphism. Since $\cos^2\alpha + \sin^2\alpha = 1$ it follows that

$\mathbf{Im}(f_4) = \{\cos\alpha + i\sin\alpha | \alpha \in \mathbb{R}\} = \mathbb{T}_1$ and $\mathbf{Ker}(f_4) = \{2\pi n | n \in \mathbb{Z}\} = \langle 2\pi \rangle$. Theorem 3.3.9 shows that $\mathbb{T}_1 \cong \mathbb{R}/\langle 2\pi \rangle$.

We modify this example a little bit further and consider the mapping $f_5 : \mathbb{R} \longrightarrow \mathbb{C}^*$ defined by $f_5(\alpha) = \cos(2\pi\alpha) + i\sin(2\pi\alpha)$, for each $\alpha \in \mathbb{R}$.

We have, since $\cos\gamma + i\sin\gamma = e^{i\gamma}$ for all γ,

$$\begin{aligned} f_5(\alpha+\beta) &= \cos(2\pi(\alpha+\beta)) + i\sin(2\pi(\alpha+\beta)) \\ &= \cos(2\pi\alpha + 2\pi\beta) + i\sin(2\pi\alpha + 2\pi\beta) \\ &= (\cos 2\pi\alpha + i\sin 2\pi\alpha)(\cos 2\pi\beta + i\sin 2\pi\beta) = f_5(\alpha)\cdot f_5(\beta), \end{aligned}$$

so f_5 is a homomorphism. Again $\mathbf{Im}(f_5) = \mathbb{T}_1$, but in this case $\mathbf{Ker}(f_5) = \mathbb{Z}$. Using Theorem 3.3.9 we deduce that $\mathbb{T}_1 \cong \mathbb{R}/\mathbb{Z}$.

Suppose that $\alpha \in \mathbb{Q}$, so $\alpha = \frac{m}{n}$ where $m, n \in \mathbb{Z}$ and $n \neq 0$. Then

$$f_5\left(\frac{m}{n}\right) = \cos\left(2\pi\left(\frac{m}{n}\right)\right) + i\sin\left(2\pi\left(\frac{m}{n}\right)\right).$$

We have, by de Moivre's Theorem,

$$\begin{aligned} \left(\cos\left(2\pi\left(\frac{m}{n}\right)\right) + i\sin\left(2\pi\left(\frac{m}{n}\right)\right)\right)^n &= \cos\left(2\pi\left(\frac{m}{n}\right)n\right) + i\sin\left(2\pi\left(\frac{m}{n}\right)n\right) \\ &= \cos(2\pi m) + i\sin(2\pi m) = 1. \end{aligned}$$

This shows that $f_5(\mathbb{Q}) = \mathbb{T}$, and hence $\mathbb{T} \cong \mathbb{Q}/\mathbb{Z}$.

5. Finally we define the mapping $f_6 : \mathbf{GL}_n(\mathbb{R}) \longrightarrow \mathbb{R}^*$ by $f_6(A) = \mathbf{det}(A)$ for each matrix $A \in \mathbf{GL}_n(\mathbb{R})$. Since $\mathbf{det}(AB) = \mathbf{det}(A)\mathbf{det}(B)$, f_6 is a homomorphism. Moreover, it is an epimorphism because, for each $\alpha \in \mathbb{R}^*$,

$$\begin{pmatrix} \alpha & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & 0 \end{pmatrix} \overset{f_6}{\longmapsto} \alpha.$$

Furthermore,

$$\mathbf{Ker}(f) = \{A \in \mathbf{GL}_n(\mathbb{R}) \mid \mathbf{det}(A) = 1\} = \mathbf{SL}_n(\mathbb{R}).$$

This shows independently that $\mathbf{SL}_n(\mathbb{R})$ is normal in $\mathbf{GL}_n(\mathbb{R})$ and Theorem 3.3.8 shows that $\mathbf{GL}_n(\mathbb{R})/\mathbf{SL}_n(\mathbb{R}) \cong \mathbb{R}^*$.

Here we have only given some very basic definitions and examples, appropriate for a first course in group theory and the reader can explore this fascinating topic in a variety of excellent books.

Exercise Set 3.3

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

3.3.1. Prove that every factor group of a cyclic group is likewise cyclic.

3.3.2. Let $G = \langle g \rangle$ be an infinite cyclic group and N be a proper normal subgroup of G. Prove that the factor group G/N is a finite cyclic group.

3.3.3. Define the mapping: $\Theta : \mathbb{N} \longrightarrow \{0,1\}$ by the rule $\Theta(x) = \begin{cases} 0, \text{ if } x \text{ is even,} \\ 1, \text{ if } x \text{ is odd.} \end{cases}$

Prove that Θ has the property that $\Theta(xy) = \Theta(x)\Theta(y)$. (Note: Θ is not a homomorphism since the structures involved are not groups.)

3.3.4. Prove that any factor group of an abelian group is abelian.

3.3.5. If N and M are groups with relatively prime orders, then the only homomorphism from N to M is the trivial one.

3.3.6. Let G be a simple group. Show that if $\alpha : G \to H$ is a homomorphism, then either α is trivial or H has a subgroup isomorphic with G.

3.3.7. Let $H \leq G$ and $N \lhd G$. Prove that $HN = \{hn | h \in H, n \in N\}$ is a subgroup of G.

3.3.8. Let $H \lhd K \leq G$ and $L \lhd G$. Show that $HL/L \lhd KL/L$.

3.3.9. A subgroup N is a maximal normal subgroup of the group G if $N \lhd G$ and there exists no normal subgroup strictly between N and G. Prove that G/N has no proper normal subgroups. Is the converse statement correct?

3.3.10. Let G be an abelian group and let N denote the set of elements in G of finite order. Prove that N is a subgroup of G and that the only element of G/N that is of finite order is the identity element.

3.3.11. Let G be a group and let $N \lhd G$. Let H be a subgroup of G. Prove the Second Isomorphism Theorem that $HN/N \cong H/N \cap N$.

3.3.12. Let G, H be groups and let $f : G \longrightarrow H$ be an epimorphism. Prove that if $N \lhd G$ then $f(N) \lhd H$.

3.3.13. Prove the Third Isomorphism Theorem, that if $N, M \triangleleft G$ for some group G and if $N \leq M$ then $(G/N)/(M/N) \cong G/M$.

3.3.14. Let G be a group and let $Z(G)$ denote the centre of G. Prove that if $G/Z(G)$ is cyclic then G is already abelian.

3.3.15. Let N be a normal subgroup of a finite group G and suppose that $\mathbf{GCD}(|N|, |G:N|) = 1$. Prove that if $|N| = k$ and $x \in G$ then the equation $x^k = e$ implies that $x \in N$.

3.3.16. Let G be an abelian group and let k be a fixed natural number. Prove that the map $f : G \longrightarrow G$ defined by $f(x) = x^k$ is a homomorphism and find its kernel.

3.3.17. Let N, M be normal subgroups of the group G and suppose that $N \cap M = \{e\}$. Prove that for all $x \in N$, for all $y \in M$ we have $xy = yx$. Give an example to show that this is not true if one of M, N is not normal.

3.3.18. There are five nonisomorphic groups of order 8. Find a concrete representation of each one and show that these are nonisomorphic.

3.3.19. Let $\mathbb{Z}[X]$ denote the group of all polynomials with integer coefficients, under addition and let p_i be the i-th prime. Prove that the function $f : \mathbb{Z}[X] \longrightarrow \mathbb{Q}^+$ defined by $f(a_0 + a_1 x + \ldots a_n x^N) = 2^{a_0} \cdot 3^{a_1} \cdot \cdots \cdot p_n^{a_n}$ is a an isomorphism.

3.3.20. Give an example of a group G having normal subgroups N, M such that $N \cong M$ but G/N is not isomorphic to G/M.

4

RINGS

4.1 RINGS, SUBRINGS, ASSOCIATIVE RINGS

In Chapter 3 we considered an important type of set with a single binary algebraic operation. Another common type of example, with strong connections to the situation of Chapter 3, is often studied, that of a set with two or more binary operations. Examples of this new situation typically involve sets of numbers, sets of polynomials, the sets of real and complex functions, the set of matrices, and the set of vectors. It was realized many years ago that these sets exhibit many common features and these properties form the basis of the definition of a ring. The concept of a ring was introduced by Richard Dedekind. But the term "ring" first appears in those works of David Hilbert dedicated to number rings. The axiomatic approach to the notion of a ring was given by Adolph Fraenkel and Emmy Noether.

Definition 4.1.1. *A set R together with two binary algebraic operations, addition and multiplication, usually denoted by $+$ and $\cdot$, is called a ring if the following conditions hold:*

(R 1) *the operation of addition has the properties*

(a) *addition is commutative, so $a+b=b+a$ for all $a,b \in R$;*

An Introduction to Essential Algebraic Structures, First Edition. Martyn R. Dixon, Leonid A. Kurdachenko and Igor Ya. Subbotin.

(b) *addition is associative, so* $a+(b+c)=(a+b)+c$, *for all* $a,b,c,\in R$;

(c) *there exists an element* $0_R \in R$ *such that* $a+0_R=a$, *for every* $a \in R$;

(d) *each element* $a \in R$ *has an additive inverse* $-a \in R$, *called an opposite, or negative, such that* $a+(-a)=0_R$.

(R 2) *addition and multiplication are connected by the distributive laws:* $(a+b)c=ac+bc$ *and* $a(b+c)=ab+ac$ *for all* $a,b,c \in R$.

Thus the elements of a ring form an abelian group under the operation of addition and we will call R, together with just the operation of addition, the additive group of R. The element 0_R, uniquely determined because of the results of Chapter 3, is usually called the *zero element* of R and when there is no ambiguity it is often simply denoted by 0. The existence of negative elements allows us to introduce the operation of subtraction in R by making the definition that $a-b=a+(-b)$.

There are a number of elementary consequences that follow from the definition of a ring, which we now list.

Proposition 4.1.2. *Let* R *be a ring and let* a,b,c *be elements of* R. *Then the following properties hold:*

(i) $a \cdot 0_R = 0_R \cdot a = 0_R$;

(ii) $a(-b)=(-a)b=-ab$;

(iii) $a(b-c)=ab-ac$ *and* $(a-b)c=ac-bc$.

In particular $(-a)(-b)=ab$.

Proof. For each $b \in R$ we have $b+0_R=b$. By the distributive law,

$$ab=a\cdot(b+0_R)=ab+a\cdot 0_R.$$

Since the element $ab \in R$ has a negative, $-ab$, we can add it to both sides and the equality becomes

$$-ab+ab=-ab+ab+a\cdot 0_R.$$

Thus $0_R=a\cdot 0_R$, since $-ab+ab=0_R$ and, likewise, $0_R\cdot a=0_R$, by a similar argument. Therefore (i) follows.

To prove (ii) we proceed as follows. From the definition of the negative element and the distributive laws we obtain

$$0_R=a\cdot 0_R=a(b+(-b))=ab+a(-b) \text{ and}$$
$$0_R=0_R\cdot b=(a+(-a))\cdot b=ab+(-a)\cdot b.$$

These equations show that $a(-b)$ is the negative of ab and $(-a)b$ is also the negative of ab and hence (ii) follows. Also by replacing a by $-a$ we see that $(-a)(-b) = (-(-a))b = ab$ since the negative of $-a$ is a itself.

These equations show that subtraction and multiplication are also connected by the distributive laws since

$$a(b-c) = a(b+(-c)) = ab + a(-c) = ab - ac$$

and likewise $(a-b)c = ac - bc$. This proves (iii) and completes the proof.

As with groups we consider certain subsets of rings that retain the structure of the ring.

Definition 4.1.3. *Let R be a ring. A subset H of R is called a subring, if H is closed under the operations of addition and multiplication and H is also a ring under the restrictions of these operations to the set H. When H is a subring of R we shall write $H \leq R$.*

We next give a criterion for a nonempty subset of a ring to be a subring. It is very similar to the criterion for a subset of a group to be a subgroup.

Theorem 4.1.4. *Let R be a ring. If H is a subring of R, then H satisfies the following conditions:*

(SR 1) *if $x, y \in H$, then $x - y \in H$;*
(SR 2) *if $x, y \in H$, then $xy \in H$.*

*Conversely, suppose that H is a nonempty subset of R. If H satisfies the conditions **(SR 1)** and **(SR 2)**, then H is a subring of R.*

Proof. Let H be a subring of R. Certainly H is a closed subset under addition and multiplication and it also follows that H has a zero element 0_H. Thus $x + 0_H = x$ for each element $x \in H$. By Definition 4.1.1, there is an element $-x \in R$ and we have $-x + x + 0_H = -x + x$ so that $0_R + 0_H = 0_R$. It follows that $0_H = 0_R$. By Definition 4.1.1, for each element $x \in H$, there is an element $y \in H$ such that $x + y = 0_H$ and since $0_H = 0_R$ it follows that y is a negative of x. As we saw in Section 3.1, each negative element is uniquely determined and hence $y = -x$. In particular, $-x \in H$. Now if x, y are arbitrary elements of H then $-y \in H$ and, since H is a closed subset under addition, we have

$$x - y = x + (-y) \in H.$$

Hence H satisfies the condition **(SR 1)**.

Since H is a closed subset under multiplication, $xy \in H$, so that H satisfies **(SR 2).**

Conversely, suppose that H is not empty and satisfies **(SR 1)** and **(SR 2)**. If $u \in H$ then, by **(SR 1)**, $0_R = u - u \in H$. If x is an arbitrary element of H then, by **(SR 1)**, $-x = 0_R - x \in H$. If also y is an arbitrary element of H then $-y \in H$ and, using **(SR 1)**, we obtain $x + y = x - (-y) \in H$. Hence H is a closed subset under addition. The condition **(SR 2)** shows that H is a closed subset under multiplication. Thus the restrictions of addition and multiplication to H are binary operations on H. The other conditions of **(R 1)** and **(R 2)** are valid for H because they are valid for all elements of R. This completes the proof.

If R is a ring, then the subsets $\{0_R\}$ and R are always subrings of R. Thus every nonzero ring has at least two subrings.

Corollary 4.1.5. *Let R be a ring and let $\mathfrak{S}$ be a family of subrings of R. The intersection $\cap\mathfrak{S}$ of all subrings of this family is also a subring in R.*

Proof. Let $S = \cap\mathfrak{S}$. We verify **(SR 1)** and **(SR 2)**. Since $0_R \in U$ for all $U \in \mathfrak{S}$ it follows that $0_R \in S \neq \emptyset$. Let $x, y \in S$. Then $x - y, xy \in U$, for all subrings $U \in \mathfrak{S}$ and therefore $x - y, xy$ belongs to the intersection of all such subrings. Thus $x - y, xy \in S$ and Theorem 4.1.4 implies that S is a subring of R.

We note that a union of subrings is not necessarily a subring. A little later we will consider a simple example confirming this.

Definition 4.1.6. *Let R be a ring.*

(R 3) *R is called associative if the multiplication in R is associative so $a(bc) = (ab)c$ for all $a, b, c \in R$;*

(R 4) *R is a ring with identity if R has an identity element e relative to the operation of multiplication so $ae = ea = a$ for all $a \in R$;*

(R 5) *R is called commutative, if the multiplication in R is commutative so $ab = ba$ for all $a, b \in R$.*

We shall often denote the element e by 1_R or 1. We note that neither 0_R nor 1_R need have any connection with the integers $0, 1$. Next we suppose that R is a ring for which $0_R = 1_R$. Then we have, using Proposition 4.1.2,

$$a = a1_R = a0_R = 0_R$$

for each $a \in R$. Hence, if the zero element of a ring R coincides with its identity element, then R consists only of the zero element. Therefore we will consider only rings whose zero element and identity element are distinct.

If R is an associative ring, then it is possible to show that it is a subring of a ring with identity. For this reason, in future, we will often consider only associative rings with identity. **Thus, from this point on we shall usually consider a ring to be an associative ring with identity whose zero element and identity element are distinct**. Thus to show that a structure is a ring we will often need to show that **(R1)–(R4)** hold, and we now take these properties to be our working definition of a ring. Sometimes we will emphasize that our rings contain an identity. We note however that this understanding will not apply to subrings. Thus we will understand the term "subring" to mean a nonempty subset S of a ring R satisfying the properties of Theorem 4.1.4 and we will not demand that a subring necessarily contains the identity. Of course multiplication will still be associative in S, since it is associative in R. Some authors make a distinction between associative and nonassociative rings. We also note that in some of our exercises the assumption that a ring always contain a multiplicative identity will not be adhered to.

We shall call a subring H of a ring R *unitary* if *H contains the identity of the ring R*. For the ring R we let

$$\begin{aligned}\mathbf{U}(R) &= \{x \in R \mid x \text{ is an invertible element of } R\}\\ &= \{x \in R | xy = yx = 1, \text{ for some } y \in R\}.\end{aligned}$$

We remark that the subset $\mathbf{U}(R)$ is closed under multiplication. Indeed, let $x, y \in \mathbf{U}(R)$. Then $y^{-1}x^{-1}xy = 1 = xyy^{-1}x^{-1}$, so that the element $y^{-1}x^{-1}$ is a multiplicative inverse of xy, and hence $xy \in \mathbf{U}(R)$. Hence restricting the multiplication of R to $\mathbf{U}(R)$ gives a binary operation on $\mathbf{U}(R)$. The multiplication on $\mathbf{U}(R)$ is associative, because it is associative on R. Furthermore, $1 \in \mathbf{U}(R)$ and if $x \in \mathbf{U}(R)$, then $(x^{-1})^{-1} = x$ so that $x^{-1} \in \mathbf{U}(R)$. Thus $\mathbf{U}(R)$ satisfies the conditions of Definition 3.1.1, so $\mathbf{U}(R)$ is a group under multiplication. The group $\mathbf{U}(R)$ is called the group of invertible elements, or the multiplicative group, of R. Since $0_R \neq 1_R, 0_R \notin \mathbf{U}(R)$.

Definition 4.1.7. *A ring R is called a division ring if every nonzero element of R has a multiplicative inverse. In this case, $\mathbf{U}(R) = R \setminus \{0_R\}$. A commutative division ring is called a field, and a noncommutative division ring is sometimes called a skew field.*

Thus in a field, the set R is an abelian group under addition and the set $R \setminus \{0_R\}$ is an abelian group under multiplication. We note also that if $ab = 0_R$ and $a \in \mathbf{U}(R)$. Then

$$0_R = a^{-1} \cdot 0_R = a^{-1}(ab) = (a^{-1}a)b = 1_R \cdot b = b.$$

However if the element a is not invertible, then a completely different situation can arise.

Definition 4.1.8. *A nonzero element a of a ring R is called a left (respectively right) zero-divisor, if there is a nonzero element b such that $ab = 0_R$ (respectively $ba = 0_R$).*

For commutative rings, if $ab = 0_R$ then $ba = 0_R$, so every left zero-divisor is a right zero-divisor and conversely. An element that is both a left zero-divisor and a right zero-divisor is simply called a zero-divisor. As we saw above, an invertible element cannot be a zero-divisor. It follows that a division ring or a field contains no zero-divisors.

Proposition 4.1.9. *Let R be a ring and let $a, x, y \in R$. Suppose that $a \neq 0$.*

(i) *If a is not a left zero-divisor and if $ax = ay$ then $x = y$;*
(ii) *If a is not a right zero-divisor and if $xa = ya$ then $x = y$.*

Proof. Since the proofs of (i) and (ii) are similar we merely prove (i). If $ax = ay$, it follows that $ax - ay = 0_R$. By the distributive law, we obtain $a(x-y) = 0_R$. Since a is not a left zero-divisor, this implies that $x - y = 0_R$ and hence $x = y$.

Proposition 4.1.9 is called the left (or right) cancellation law. We obtain the following immediate consequence.

Corollary 4.1.10. *Let R be a ring with no zero-divisors. If $a, x, y \in R$, where $r \neq 0$, and $ax = ay$ (respectively $xa = ya$) then $x = y$.*

The set of integers has some very standard properties that are common to certain types of rings which we now introduce.

Definition 4.1.11. *A ring R is called an integral domain if R is commutative, has a multiplicative identity and has no zero-divisors.*

Thus, as we saw above, every field is an integral domain. The ring $\mathbb{Z}$ is an example of an integral domain that is not a field. We give next some very natural examples of rings.

1. The set $\mathbb{C}$ of all complex numbers, the set $\mathbb{R}$ of all real numbers and the set $\mathbb{Q}$ of all rational numbers are fields. Moreover, $\mathbb{Q}$ is a (unitary) subring of $\mathbb{R}$, $\mathbb{R}$ is a (unitary) subring of $\mathbb{C}$, and the set $\mathbb{Z}$ of all integers is a unitary subring of $\mathbb{Q}$. As we saw in Section 3.1, if n is a fixed natural number or zero, the subset $n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}$ is a subgroup of the additive group of $\mathbb{Z}$ for each $n \neq 0$. In particular, it satisfies the condition (**SR 1**). Clearly $n\mathbb{Z}$ satisfies (**SR 2**). Using Theorem 4.1.4, we see that $n\mathbb{Z}$ is a subring of $\mathbb{Z}$. We remark that this subring does not contain an identity element if $n \neq 1$.

Conversely, let H be a subring of $\mathbb{Z}$. Then H satisfies the condition (**SR 1**). However condition (**SR 1**) is just condition (**ASG 3**) in disguise. Using Corollary 3.1.5, we deduce that H is a subgroup of the additive group $\mathbb{Z}$ and we have already seen in this case that $H = n\mathbb{Z}$ for some integer $n \geq 0$.

In particular, $2\mathbb{Z}$ and $3\mathbb{Z}$ are subrings, but their union does not contain the integer $5 = 2+3$. Hence $2\mathbb{Z} \cup 3\mathbb{Z}$ is not a subring, which illustrates that, in general, a union of two subrings need not be a subring.

Next, let p be a prime and let

$$\mathbb{Q}_p = \left\{ \frac{m}{p^k} \mid m,k \in \mathbb{Z} \right\}.$$

As we have already observed in Section 3.1, $\mathbb{Q}_p$ is a subgroup of the additive group $\mathbb{Q}$. In particular, it satisfies the condition (**SR 1**). Clearly $\mathbb{Q}_p$ also satisfies (**SR 2**), since

$$\frac{m}{p^k} \cdot \frac{s}{p^l} = \frac{ms}{p^{k+l}} \text{ for all } \frac{m}{p^k}, \frac{s}{p^l} \in \mathbb{Q}_p.$$

Using Theorem 4.1.4, we deduce that $\mathbb{Q}_p$ is a subring of $\mathbb{Q}$, which is also unitary, called the ring of p-adic fractions.

There are many important examples of subrings of $\mathbb{C}$ and we now consider one such. Let r be an integer with the property $\sqrt{r} \notin \mathbb{Q}$ and let

$$\mathbb{Z}[\sqrt{r}] = \{a + b\sqrt{r} \mid a,b \in \mathbb{Z}\} \neq \emptyset.$$

If α, β are arbitrary elements of $\mathbb{Z}[\sqrt{r}]$, say $\alpha = a + b\sqrt{r}$, $\beta = a_1 + b_1\sqrt{r}$, where $a, a_1, b, b_1 \in \mathbb{Z}$, then

$$\alpha - \beta = (a - a_1) + (b - b_1)\sqrt{r}, \text{ and } \alpha\beta = (aa_1 + bb_1 r) + (ab_1 + ba_1)\sqrt{r}.$$

Thus $\alpha - \beta, \alpha\beta \in \mathbb{Z}[\sqrt{r}]$. By Theorem 4.1.4, $\mathbb{Z}[\sqrt{r}]$ is a subring of $\mathbb{C}$. Clearly $1 \in \mathbb{Z}[\sqrt{r}]$ and we call $\mathbb{Z}[\sqrt{r}]$ a quadratic extension of the ring $\mathbb{Z}$.

2. In what follows we use the fact that two functions with the same domain and codomain are equal precisely when they both take the same value at every point in the domain of the functions. We shall also use many properties of the real numbers, such as the associativity of addition.

Let M be an arbitrary subset of $\mathbb{R}$ and consider the set $\mathbb{R}^M$ of all functions $f : M \longrightarrow \mathbb{R}$. We define the sum, $f + g$, and the product, $f \cdot g = fg$ of two such functions f, g as usual by

$$(f + g)(a) = f(a) + g(a), (f \cdot g)(a) = f(a)g(a)$$

for every element $a \in M$. We note here that the product is not defined via composition of mappings.

If $f, g \in \mathbb{R}^M$, then

$$(f+g)(a) = f(a)+g(a) = g(a)+f(a) = (g+f)(a)$$

for each element $a \in M$. It follows that $f+g = g+f$. Similarly, for f, g, $h \in \mathbb{R}^M$ we have

$$\begin{aligned}(f+(g+h))(a) &= f(a)+(g+h)(a) = f(a)+(g(a)+h(a))\\ &= (f(a)+g(a))+h(a) = (f+g)(a)+h(a)\\ &= ((f+g)+h)(a),\end{aligned}$$

for an arbitrary element $a \in M$, so that $f+(g+h) = (f+g)+h$.

We define a mapping $\theta : M \longrightarrow \mathbb{R}$ by $\theta(a) = 0$ for each $a \in M$. Then

$$(f+\theta)(a) = f(a)+\theta(a) = f(a)+0 = f(a), \text{ so } f+\theta = f,$$

for every $f \in \mathbb{R}^M$. This means that θ is the zero element of $\mathbb{R}^M$. Finally, we define the mapping $-f \in \mathbb{R}^M$ by the rule $(-f)(a) = -f(a)$ and it is easy to see that $f+(-f) = \theta$.

Next we let $f, g, h \in \mathbb{R}^M$, $a \in M$. Then

$$\begin{aligned}f \cdot (g+h)(a) &= f(a) \cdot (g+h)(a) = f(a)(g(a)+h(a))\\ &= f(a)g(a)+f(a)h(a) = (f \cdot g)(a)+(f \cdot h)(a)\\ &= (f \cdot g+f \cdot h)(a),\end{aligned}$$

so $f \cdot (g+h) = f \cdot g+f \cdot h$. In a similar way we can show that $(f+g) \cdot h = f \cdot h+g \cdot h$.

Multiplication of functions is also associative as the following argument shows:

$$\begin{aligned}f \cdot (gh)(a) &= f(a) \cdot (gh)(a) = f(a)(g(a)h(a)) = (f(a)g(a))h(a)\\ &= (fg)(a)h(a) = (fg) \cdot h(a).\end{aligned}$$

It is easy to see, using a similar method, that multiplication is commutative. The identity element is the function $I \in \mathbb{R}^M$, defined by $I(a) = 1$ for each element $a \in M$ and we have:

$$(fI)(a) = f(a)I(a) = f(a) \cdot 1 = f(a), \text{ so } fI = f.$$

Thus $\mathbb{R}^M$ satisfies the conditions (**R 1**)–(**R 5**). Hence $\mathbb{R}^M$ is a commutative ring.

The multiplicative group, $\mathbf{U}(\mathbb{R}^M)$, of $\mathbb{R}^M$ is easy to describe. The reader will readily verify that $\mathbf{U}(\mathbb{R}^M)$ consists of the subset of all functions f such that $f(a) \neq 0$ for all $a \in M$. In this case the inverse g of f is the function defined by $g(a) = \big(f(a)\big)^{-1}$, for all $a \in M$; in order to avoid possible confusion with certain other notation we shall not denote the multiplicative inverse of f here by f^{-1} (this notation is usually reserved for the inverse of a bijection, for instance).

We note that the ring $\mathbb{R}^M$ has zero-divisors, provided M has at least two elements. For example, let L be a nonempty subset of M such that $M \setminus L$ is also nonempty, and define functions f, g by

$$f(a) = \begin{cases} 1 \text{ if } a \in L \\ 0 \text{ if } a \notin L, \end{cases} \qquad \text{and } g(a) = \begin{cases} 0 \text{ if } a \in L \\ 1 \text{ if } a \notin L. \end{cases}$$

The functions f, g are nonzero, but $f(a)g(a) = 0$ for every $a \in M$ and hence $fg = \theta$.

In Calculus courses we usually deal with the situation when the set M is either $\mathbb{R}$ or an interval $[a,b] \subseteq \mathbb{R}$. The ring $\mathbb{R}^{\mathbb{R}}$ has numerous well-known unitary subrings, including the subring of all continuous functions, the subring of all differentiable functions, the subring of all bounded functions, and so on.

3. Next we consider the set $\mathbf{M}_n(\mathbb{R})$ of all square matrices of order n, whose coefficients belong to the field $\mathbb{R}$ of all real numbers. It is quite easy to see that under the usual operations of matrix addition and multiplication $\mathbf{M}_n(\mathbb{R})$ is a ring. Furthermore, a matrix A is invertible if and only if $\mathbf{det}(A) \neq 0$. This means that $\mathbf{U}(\mathbf{M}_n(\mathbb{R})) = \mathbf{GL}_n(\mathbb{R})$. The ring $\mathbf{M}_n(\mathbb{R})$ is not commutative if $n \geq 2$, and it contains zero-divisors; for example, $E_{jm}E_{rt} = 0$, when $m \neq r$. Here E_{jm} is a basic matrix so $E_{jm} = [u_{ks}] \in \mathbf{M}_n(\mathbb{R})$, where $u_{jm} = 1$ and $u_{ks} = 0$ whenever $(k,s) \neq (j,m)$. Thus for example in $\mathbf{M}_2(\mathbb{R})$, $E_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $E_{12}E_{12}$ is the zero matrix.

We let $\mathbf{T}_n(\mathbb{R})$ denote the subset of $\mathbf{M}_n(\mathbb{R})$, consisting of all (upper) triangular matrices, not just the nonsingular ones. Clearly $\mathbf{T}_n(\mathbb{R})$ satisfies the condition (**SR 1**). Also, if $A = [a_{jt}], B = [b_{jt}] \in \mathbf{T}_n(\mathbb{R})$ and $C = AB = [c_{jt}]$, then

$$c_{jt} = a_{j,1}b_{1,t} + a_{j,2}b_{2,t} + \cdots + a_{j,j-1}b_{j-1,t} + a_{jj}b_{jt} + a_{j,j+1}b_{j+1,t} + \cdots + a_{jn}b_{nt}.$$

If $j > t$ then, since

$$a_{j,1} = a_{j,2} = \cdots = a_{j,j-1} = 0 \text{ and } b_{jt} = b_{j+1,t} = \cdots = b_{nt} = 0,$$

we have also $c_{jt} = 0$. It follows that $C = AB \in \mathbf{T}_n(\mathbb{R})$. Thus $\mathbf{T}_n(\mathbb{R})$ also satisfies the condition (**SR 2**), and Theorem 4.1.4 shows that $\mathbf{T}_n(\mathbb{R})$ is a subring of $\mathbf{M}_n(\mathbb{R})$. Clearly, this subring is unitary. We observe that

$$\begin{aligned} c_{jj} &= a_{j,1}b_{1,j} + a_{j,2}b_{2,j} + \cdots + a_{jj}b_{jj} + a_{j,j+1}b_{j+1,j} + \cdots + a_{jn}b_{nj} \\ &= a_{jj}b_{jj}. \end{aligned}$$

A triangular matrix $A = [a_{jt}] \in \mathbf{M}_n(\mathbb{R})$ is called *zero-triangular*, if $a_{11} = a_{22} = \cdots = a_{nn} = 0$ and we denote the set of zero-triangular matrices by $\mathbf{NT}_n(\mathbb{R})$. As above, it is easy to see that $\mathbf{NT}_n(\mathbb{R})$ satisfies the conditions (**SR 1**) and (**SR 2**), and by Theorem 4.1.4, it is a subring. This subring does not contain the identity element of $\mathbf{M}_n(\mathbb{R})$ and it contains no invertible elements of $\mathbf{M}_n(\mathbb{R})$.

Let $\mathbf{D}_n(\mathbb{R})$ denote the subset of $\mathbf{T}_n(\mathbb{R})$, consisting of all diagonal matrices and let $\mathbb{R}I = \{\lambda I \mid \lambda \in \mathbb{R}\}$, the subset of all scalar matrices (where I is the $n \times n$ identity matrix). Theorem 4.1.4 again shows that $\mathbf{D}_n(\mathbb{R})$ and $\mathbb{R}I$ are unitary subrings of $\mathbf{M}_n(\mathbb{R})$. Finally, let

$$\mathbb{R}E_{ii} = \{\lambda E_{ii} \mid \lambda \in \mathbb{R}\}.$$

We have

$$\lambda E_{ii} - \mu E_{ii} = (\lambda - \mu)E_{ii} \text{ and } \lambda E_{ii} \cdot \mu E_{ii} = \lambda\mu E_{ii}.$$

Theorem 4.1.4 implies that $\mathbb{R}E_{ii}$ is a subring of $\mathbf{M}_n(\mathbb{R})$. Furthermore,

$$\lambda E_{ii} \cdot E_{ii} = E_{ii} \cdot \lambda E_{ii} = \lambda E_{ii}.$$

This equation shows that E_{ii} is the multiplicative identity element of the subring $\mathbb{R}E_{ii}$, so $\mathbb{R}E_{ii}$ is not a unitary subring of $\mathbf{M}_n(\mathbb{R})$, but nevertheless it has its own multiplicative identity element. Thus a nonunitary subring of a ring can have a different multiplicative identity element from that of the ring. Furthermore, if $\lambda \neq 0$ then λE_{ii} has an inverse in $\mathbb{R}E_{ii}$, namely $\lambda^{-1}E_{ii}$, even though λE_{ii} is not invertible as an element of the ring $\mathbf{M}_n(\mathbb{R})$.

Using virtually the same reasoning, one can show that the set $\mathbf{M}_n(\mathbb{C})$ of all matrices of order n with complex coefficients is also a ring. Furthermore, $\mathbf{M}_n(\mathbb{R})$ is a unitary subring of $\mathbf{M}_n(\mathbb{C})$. Moreover, let H be a subring

of $\mathbb{C}$ and consider the subset $\mathbf{M}_n(H)$ of $\mathbf{M}_n(\mathbb{C})$ consisting of all matrices whose coefficients belong to H. Let $A = [a_{jt}], B = [b_{jt}] \in \mathbf{M}_n(H)$, and let $C = A - B = [c_{jt}]$, $D = AB = [d_{jt}]$. Then we have

$$c_{jt} = a_{jt} - b_{jt} \in H, \text{ for } 1 \leq j,t \leq n,$$
$$d_{jt} = a_{j,1}b_{1,t} + a_{j,2}b_{2,t} + \cdots + a_{jn}b_{nt} \in H.$$

Hence $\mathbf{M}_n(H)$ satisfies the conditions (**SR 1**), (**SR 2**) and, by Theorem 4.1.4, it is a subring. If H is a unitary subring of $\mathbb{C}$, then clearly $\mathbf{M}_n(H)$ is a unitary subring of $\mathbf{M}_n(\mathbb{C})$. Thus $\mathbf{M}_n(\mathbb{Q})$ and $\mathbf{M}_n(\mathbb{Z})$ are unitary subrings of $\mathbf{M}_n(\mathbb{C})$.

4. In the ring $\mathbf{M}_4(\mathbb{R})$ we consider now the following very specific example. In Section 3.2 we introduced the following matrices in the group $\mathbf{GL}_4(\mathbb{R})$, albeit using a different notation.

$$I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$
$$B = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}.$$

The multiplication table for these matrices was given in Section 3.2, but we reproduce part of it here, using the current notation:

	I	A	B	C
I	I	A	B	C
A	A	$-I$	C	$-B$
B	B	$-C$	$-I$	A
C	C	B	$-A$	$-I$

Now let $\mathbf{H} = \{x_1I + x_2A + x_3B + x_4C \mid x_1, x_2, x_3, x_4 \in \mathbb{R}\}$. Let

$$\alpha = x_1I + x_2A + x_3B + x_4C, \beta = y_1I + y_2A + y_3B + y_4C \in \mathbf{H},$$

the coefficients of course coming from $\mathbb{R}$. We have

$$\alpha - \beta = (x_1 - y_1)I + (x_2 - y_2)A + (x_3 - y_3)B + (x_4 - y_4)C \in \mathbf{H},$$

and, using the multiplication table,

$$\begin{aligned}\alpha\beta &= (x_1I + x_2A + x_3B + x_4C)(y_1I + y_2A + y_3B + y_4C)\\ &= (x_1y_1 - x_2y_2 - x_3y_3 - x_4y_4)I + (x_1y_2 + x_2y_1 + x_3y_4 - x_4y_3)A\\ &\quad + (x_1y_3 - x_2y_4 + x_3y_1 + x_4y_2)B + (x_1y_4 + x_2y_3 - x_3y_2 + x_4y_1)C\\ &\quad \in \mathbf{H}.\end{aligned}$$

Thus we see that $\mathbf{H}$ satisfies the conditions (**SR 1**), (**SR 2**) and, by Theorem 4.1.4, it is a subring of $\mathbf{M}_4(\mathbb{R})$. The equation $I = 1I + 0A + 0B + 0C$ shows that $\mathbf{H}$ is a unitary subring and clearly $\mathbf{H}$ is not commutative. Next we determine the group $\mathbf{U}(\mathbf{H})$. Let $\alpha = x_1I + x_2A + x_3B + x_4C \neq 0$ so that $(0,0,0,0) \neq (x_1,x_2,x_3,x_4)$. We consider the product

$$\begin{aligned}&(x_1I + x_2A + x_3B + x_4C)(x_1I - x_2A - x_3B - x_4C)\\ &\quad = (x_1x_1 + x_2x_2 + x_3x_3 + x_4x_4)I + (-x_1x_2 + x_2x_1 - x_3x_4 + x_4x_3)A\\ &\qquad + (-x_1x_3 + x_2x_4 + x_3x_1 - x_4x_2)B + (-x_1x_4 - x_2x_3 + x_3x_2 + x_4x_1)C\\ &\quad = (x_1^2 + x_2^2 + x_3^2 + x_4^2)I.\end{aligned}$$

By our assumption, $d = x_1^2 + x_2^2 + x_3^2 + x_4^2$ is a positive real number. Let

$$z_1 = \frac{x_1}{d}, z_2 = -\frac{x_2}{d}, z_3 = -\frac{x_3}{d}, z_4 = -\frac{x_4}{d}, \text{ and } \beta = (z_1I + z_2A + z_3B + z_4C).$$

Then, as above, we see that $\alpha\beta = I$. Since it is also easy to show that $\beta\alpha = I$, it follows that α is invertible and $\beta = \alpha^{-1}$. Thus every nonzero element of $\mathbf{H}$ has a multiplicative inverse, which means that $\mathbf{H}$ is a division ring, but not a field. We call $\mathbf{H}$ the *Ring of Quaternions*.

5. Finally we consider the following standard subring of an arbitrary ring.

If R is a ring, then set

$$\zeta(R) = \{z \in R \mid zx = xz \text{ for all } x \in R\}.$$

Thus $\zeta(R)$ is the set of all elements of R that commute with all other elements of R. Clearly $\zeta(R) \neq \emptyset$ since $0_R \in \zeta(R)$.

Let $a, b \in \zeta(R), x \in R$. Then $ax = xa, bx = xb$. Hence

$$\begin{aligned}&(a-b)x = ax - bx = xa - xb = x(a-b), \text{ and}\\ &(ab)x = a(bx) = a(xb) = (ax)b = (xa)b = x(ab).\end{aligned}$$

These equations show that $a-b, ab \in \zeta(R)$. By Theorem 4.1.4, $\zeta(R)$ is a subring of R, called *the center of* R.

Let 1 be the identity element of R and let $\mathbb{Z}1 = \{n1 \mid n \in \mathbb{Z}\}$. We have, for $m, n \in \mathbb{Z}$,

$$m1 - n1 = (m-n)1, \text{ and } (m1)(n1) = (mn)1.$$

By Theorem 4.1.4, $\mathbb{Z}1$ is a subring of R. Since

$$(n1)a = n(1a) = na = a(n1),$$

it follows that $n1 \in \zeta(R)$. Hence $\mathbb{Z}1 \leq \zeta(R)$.

Exercise Set 4.1

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

4.1.1. On the set $R = \mathbb{Z} \times \mathbb{Z}$ we define the operations of addition and multiplication by the following rules: $(a,b) + (a_1,b_1) = (a+a_1, b+b_1)$, $(a,b)(a_1,b_1) = (0,0)$. Is R a ring under these operations? If the answer is yes, does R have an identity element or zero-divisors?

4.1.2. On the set $R = \mathbb{Z} \times \mathbb{Z}$ we define the operations of addition and multiplication by $(a,b) + (a_1,b_1) = (a+a_1, b+b_1)$, $(a,b)(a_1,b_1) = (a+a_1+b+b_1, 0)$. Is R a ring under these operations?

4.1.3. On the set $R = \mathbb{Z} \times \mathbb{Z}$ we define the operations of addition and multiplication by $(a,b) + (a_1,b_1) = (aa_1, bb_1)$, $(a,b)(a_1,b_1) = (a+a_1, b+b_1)$. Is R a ring under these operations?

4.1.4. On the set $R = \mathbb{Z} \times \mathbb{Z}$ we define the operations of addition and multiplication by the following rules: $(a,b) + (a_1,b_1) = (a+a_1, b+b_1)$, $(a,b)(a_1,b_1) = (aa_1, bb_1)$. Prove that R is a ring with identity and find its all zero-divisors. (This can be done more generally with essentially the same proof: Let R, S be rings. We can define a new ring $R \oplus S = \{(r,s) | r \in R, s \in S\}$ with operations defined by $(a,b) + (r,s) = (a+r, b+s)$ and $(a,b)(r,s) = (ar, bs)$ for all $a, r \in R, b, s \in S$.)

4.1.5. Let R be a ring, $a, b \in R$. Prove that the equation $a+x = b$ has a unique solution in R.

4.1.6. Let A be an abelian group with additive operation. If we define multiplication by $ab = a - b$, for all $a, b \in A$ then is A a ring?

4.1.7. Find a subring of $\mathbb{C}$ generated by the subset $\mathbb{Z} \cup \{\sqrt{-5}\}$. Find all invertible elements of this subring.

4.1.8. Find a subring of $\mathbb{C}$, generated by the subset $\mathbb{Z} \cup \{\sqrt{-3}\}$. Find all invertible elements of this subring.

4.1.9. Let L be the subset of $\mathbf{M}_2(\mathbb{Q})$ consisting of all matrices of the form $\begin{pmatrix} a & b \\ -b & a-b \end{pmatrix}$. Prove that L is a subring of $\mathbf{M}_2(\mathbb{Q})$. Is L commutative?

4.1.10. Give examples of zero-divisors in the ring $\mathbf{M}_2(\mathbb{Z})$.

4.1.11. Find the center of the ring $\mathbf{M}_2(\mathbb{R})$.

4.1.12. Let R be a ring. Let $\mathbf{M}_2(R) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} | a, b, c, d \in R \right\}$. Prove that $\mathbf{M}_2(R)$ with the natural operations is a ring that in general is noncommutative.

4.1.13. Let R be a ring and let I, J be subrings of R. Give an example to show that $I + J = \{i + j | i \in I, j \in J\}$ need not be a subring of R.

4.1.14. Let R be a ring with identity 1 and operations $+, \ldots$ Prove that R together with the operations $\circ, *$ is a ring if we define

$$a \circ b = a + b + 1, a * b = a + b + ab,$$

for all $a, b \in R$. Find the zero element of this ring, the negative of each element of R and the identity element.

4.1.15. Let R be a ring such that $x^2 = x$ for all $x \in R$. (Such a ring is called a Boolean ring.) Prove that R is a commutative ring.

4.1.16. Let R be a ring whose underlying additive subgroup is cyclic. Prove that R is a commutative ring.

4.1.17. Give an example of a ring with fixed elements a, b in which the equation $ax = b$ has more than one solution for x.

4.1.18. Give an example of a ring R with elements a, b, c, with $a \neq 0$, satisfying $ab = ac$, but where this does not imply that $b = c$.

4.1.19. Prove that if R is a finite integral domain then R is a field.

4.1.20. Prove that if R is an integral domain and $a, b, c \in R$, with $a \neq 0$, and such that $ab = ac$ then $b = c$.

4.2 RINGS OF POLYNOMIALS

Polynomial rings play a central role in the study of rings. Such rings are among the oldest objects of study in algebra. Polynomials are of course important in other branches of mathematics and can be used to solve a variety of problems, not only in algebra, but also in many other areas.

The study of polynomials often begins in secondary school mathematics and then later in calculus. There, polynomials are considered from a functional point of view. In other words, a polynomial function is assumed to be a function of the type $f : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$f : x \longmapsto a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,$$

where $x \in \mathbb{R}$ and (typically) the coefficients a_i are also real. With this approach we obtain the following traditional form of polynomials

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

Addition and multiplication of real functions lead to the familiar rules of addition and multiplication of polynomials.

However, this approach is not well suited to algebra at a more advanced level. First of all, this approach is connected with the question of the equality of polynomials. In calculus it is proved that two polynomials $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$ and $g(x) = b_0 + b_1 x + b_2 x^2 + \cdots + b_k x^k$ coincide (as functions) if and only if $n = k$ and $a_j = b_j$ for all j such that $0 \leq j \leq n$. In general, however, polynomials are considered whose coefficients come from different algebraic structures, and in particular, from finite fields and domains, and the functional approach is not quite so effective there. Therefore, we will think of a polynomial as a kind of formal notation. This approach can be rigorous, but our aim is not to be too formal with regard to such matters in this book. To this end we let R be a commutative (and associative) ring, with identity and let X be a symbol, commonly called a variable. An expression aX^t where $a \in R$, and t is a non-negative integer is called a monomial in X with the coefficient $a \in R$. We define $aX^0 = a$, so that the elements of the ring R are also monomials. We consider the formal expression aX^t, as a type of "picture" shown on the paper. For monomials, like terms can be added very naturally via $aX^t + bX^t = (a+b)X^t$ and multiplication is done by $(aX^t)(bX^m) = abX^{t+m}$, where $a, b \in R$.

An abstract polynomial (a "picture") consists of several monomials, connected by the sign +, so

$$f(X) = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n.$$

is called a polynomial in X, with coefficients in R. It is assumed that the order in which the monomials are written is irrelevant, that like terms can be added, and that we may insert and discard monomials with zero coefficients. As a rule, we will write a polynomial in its canonical form, one where the like terms have already been combined and the monomials are arranged in order of increasing powers of X.

We now give the definitions of the equality of polynomials and the basic operations on them.

Two polynomials are equal if in their canonical forms the coefficients at all like powers of X are equal, in other words, they have the same canonical representation. Thus if $f(X) = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$ and $g(X) = b_0 + b_1X + b_2X^2 + \cdots + b_kX^k$ then $f(X) = g(X)$ if and only if $n = k$ and, for each i, we have $a_i = b_i$.

The sum of two polynomials is a polynomial obtained by adding the like terms. So if $f(X) = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$ and $g(X) = b_0 + b_1X + b_2X^2 + \cdots + b_nX^n$ (note that some of the coefficients here may be zero) then their sum is

$$f(X) + g(X) = (a_0 + b_0) + (a_1 + b_1)X + (a_2 + b_2)X^2 + \cdots + (a_n + b_n)X^n.$$

In other words, to add two polynomials, we need to add up the coefficients of the same powers of X. Thus, addition of polynomials is reduced to the addition of relevant elements of R, and therefore it inherits the basic properties of addition in R. In particular, addition of polynomials is commutative:

$$f(X) + g(X) = g(X) + f(X)$$

and associative:

$$(f(X) + g(X)) + h(X) = f(X) + (g(X) + h(X)).$$

There is a zero element, the zero polynomial, which is simply the zero element of the ring R, considered as a polynomial, containing no monomials with nonzero coefficients. This zero polynomial is again denoted by 0_R or 0, when no ambiguity arises. Finally, an additive inverse to the polynomial

$$f(X) = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n.$$

is the polynomial

$$-f(X) = -a_0 - a_1X - a_2X^2 - \cdots - a_nX^n.$$

It is clear from the definition of addition that $-f(X)+f(X)$ is the zero polynomial.

The product of two polynomials is a polynomial obtained using a natural distributive law, multiplying every term of the first polynomial in the product by every term in the second, in the correct order. Of course the like terms are then combined. More precisely,

$$\begin{aligned} f(X)g(X) &= (a_0+a_1X+\cdots+a_nX^n)(b_0+b_1X+\cdots+b_kX^k) \\ &= c_0+c_1X+c_2X^2+\cdots+c_{n+k}X^{n+k}, \end{aligned}$$

where

$$c_j = a_0b_j+a_1b_{j-1}+\cdots+a_jb_0 = \sum_{0\le s\le j} a_sb_{j-s} = \sum_{s+t=j} a_sb_t.$$

It is easy to check that this multiplication is commutative and associative, since we assumed initially that multipication in the original ring R was commutative. Multiplication is also connected to addition by the distributive law. The set of polynomials has an identity element, simply the identity element of the ring R. We let

$$R[X] = \{\text{polynomials in } X \text{ with coefficients in } R\}$$

and the above discussion serves to show that $R[X]$ is a ring called the *polynomial ring with coefficients in R*.

Let $f(X) = a_0+a_1X+a_2X^2+\cdots+a_nX^n$, where $a_n \neq 0$. Thus the highest power of X with nonzero coefficient is n and in this case we call n the *degree* of the polynomial $f(X)$ and the coefficient of X^n is called the *leading coefficient* of the polynomial $f(X)$. We will denote the degree of a polynomial $f(X)$ by $\mathbf{deg}(f)$, or $\mathbf{deg}\, f(X)$. It is easy to prove the following, using our earlier observations.

Proposition 4.2.1. *Let R be a commutative ring, and let* $f(X)$, $g(X) \in R[X]$. *Then*

(i) $\mathbf{deg}(f+g) \le \mathbf{deg}(f)$, *and* $\mathbf{deg}(f+g) \le \mathbf{deg}(g)$;
(ii) $\mathbf{deg}(fg) \le \mathbf{deg}(f)+\mathbf{deg}(g)$.

Corollary 4.2.2. *Let R be a commutative ring. If the ring R has no zero-divisors, then* $\mathbf{deg}(fg) = \mathbf{deg}(f)+\mathbf{deg}(g)$. *In particular, in this case,* $R[X]$ *also has no zero-divisors. Moreover,* $\mathbf{U}(R[X]) = \mathbf{U}(R)$.

Proof. Let $f(X) = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$ and $g(X) = b_0 + b_1X + b_2X^2 + \cdots + b_kX^k$, where $a_n, b_k \neq 0$. Then the coefficient of X^{n+k} is a_nb_k, which must be nonzero, by hypothesis. Hence in this case $\mathbf{deg}(fg) = \mathbf{deg}(f) + \mathbf{deg}(g)$ and it is also clear that $R[X]$ has no zero-divisors. An element of $f(X) \in R[X]$ will be invertible if there is a polynomial $g(X)$ such that $f(X)g(X) = 1$. Since 1 is a polynomial of degree 0 it follows that $f(X)$ must also be of degree 0, so $f(X)$ must therefore be an element of R. The result follows.

Suppose that R is a unitary subring of a commutative ring K. Let $f(X) = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n \in R[X]$ and let $c \in K$. Then the element $f(c) = a_0 + a_1c + a_2c^2 + \cdots + a_nc^n$ of the ring K is called the value of $f(X)$ at $X = c$. It is easy to prove that if $f(X), g(X) \in R[X]$ and if $h(X) = f(X) + g(X), u(X) = f(X)g(X)$, then

$$h(c) = f(c) + g(c) \text{ and } u(c) = f(c)g(c).$$

Using Theorem 4.1.4, we deduce that $R[c] = \{f(c) \mid f(X) \in R[X]\}$ is a subring of K. We shall call an element $c \in K$ a root, or a zero, of the polynomial $f(X)$, if $f(c) = 0_R$. This definition implies that a root of a polynomial does not necessarily belong to the same ring as the coefficients. For example, the polynomial $X^2 - 2 \in \mathbb{Q}[X]$ has no rational roots; its roots are real numbers. The polynomial $X^2 + 1 \in \mathbb{Q}[X]$ has no real roots; its roots are complex numbers.

For the case when R is a field (and this case is especially important to us), the polynomial ring $R[X]$ has the following important property. It is analogous to the division algorithm in the integers.

Theorem 4.2.3. *Let F be a field, let $f(X), g(X) \in F[X]$ and suppose that $g(X)$ is a nonzero polynomial. Then there exist polynomials $q(X), r(X) \in F[X]$ such that $f(X) = q(X)g(X) + r(X)$, where either $r(X) = 0_F$ or $\mathbf{deg}(r(X)) < \mathbf{deg}(g(X))$. This representation is unique.*

Proof. Let

$$f(X) = a_0 + a_1X + \cdots + a_nX^n \text{ and } g(X) = b_0 + b_1X + \cdots + b_kX^k,$$

where $b_k \neq 0_F$. We will use induction on n to prove the theorem. If $\mathbf{deg}\, f(X) < \mathbf{deg}\, g(X)$ then put $r(X) = f(X)$ and $q(X) = 0_F$. Thus we may assume that $\mathbf{deg}\, f(X) \geq \mathbf{deg}\, g(X)$. If $\mathbf{deg}\, f(X) = 0$ then $\mathbf{deg}\, g(X) = 0$ and we set $r(X) = 0_F$, $q(X) = a_0b_0^{-1}$. Suppose now that $n > 0$ and suppose that our theorem has been proved for all polynomials of degree less than n. The polynomial $a_nb_k^{-1}X^{n-k}g(X)$ has degree n and its leading coefficient is a_n. Then the

degree of the polynomial $f(X) - a_n b_k^{-1} X^{n-k} g(X) = f_1(X)$ is less than n and, by induction,

$$f_1(X) = q_1(X)g(X) + r(X),$$

where either $r(X) = 0_F$ or $\mathbf{deg}\, r(X) < \mathbf{deg}\, g(X)$. We now have

$$f(X) = a_n b_k^{-1} X^{n-k} g(X) + f_1(X) = q(X)g(X) + r(X)$$

where

$$q(X) = q_1(X) + a_n b_k^{-1} X^{n-k}$$

and the existence of the decomposition follows.

Suppose also that

$$f(X) = q_2(X)g(X) + r_2(X),$$

where either $r_2(X) = 0_F$, or $\mathbf{deg}\, r_2(X) < \mathbf{deg}\, g(X)$. Then

$$q(X)g(X) + r(X) = q_2(X)g(X) + r_2(X).$$

and it follows that

$$g(X)(q(X) - q_2(X)) = r_2(X) - r(X).$$

The polynomial $r_2(X) - r(X)$ is either zero or its degree is less than $\mathbf{deg}\, g(X)$. On the other hand, if $q(X) - q_2(X) \neq 0_F$, then

$$\mathbf{deg}\,(g(X)(q(X) - q_2(X)) = \mathbf{deg}\, g(X) + \mathbf{deg}\,(q(X) - q_2(X)) \geq \mathbf{deg}\, g(X),$$

which gives a contradiction. Thus $q_2(X) - q(X) = 0_F$, which implies that $q(X) = q_2(X)$ and hence $r_2(X) = r(X)$. This establishes the uniqueness portion of the result and completes the proof.

We emphasize the specific case when the remainder is zero in Theorem 4.2.3.

Definition 4.2.4. *Let F be a field and let $f(X), g(X) \in F[X]$. We say that $g(X)$ divides $f(X)$ or $g(X)$ is a divisor of $f(X)$ if $f(X) = g(X)h(X)$ for some polynomial $h(X) \in F[X]$.*

Let F be a field, let $f(X) \in F[X]$ and $c \in F$. Applying Theorem 4.2.3 to the polynomials $f(X)$ and $(X-c)$ we deduce that $f(X) = q(X)(X-c)+r(X)$ where either $r(X) = 0_F$ or $\mathbf{deg}\, r(X) < \mathbf{deg}\,(X-c)$. Since $\mathbf{deg}\,(X-c) = 1$, it follows that either $r(X) = 0_F$ or $\mathbf{deg}\, r(X) = 0$. This means that $r(X) = b$ is an element of F. Thus $f(c) = q(c)(c-c)+b = b$. In particular, we obtain the factor theorem.

Corollary 4.2.5. *Let F be a field, and let $f(X) \in F[X]$. If an element c is a root of a polynomial $f(X)$, then $f(X) = q(X)(X-c)$ for some polynomial $q(X) \in F[X]$. Hence c is a zero of $f(X)$ if and only if $X-c$ is a factor of $f(X)$.*

Since every root of $q(X)$ is also a root of $f(X)$, we obtain

Corollary 4.2.6. *Let F be a field, and let $f(X) \in F[X]$. Then $f(X) = q(X)$ $(X-c_1)\dots(X-c_t)$ for some polynomial $q(X) \in F[X]$ such that $q(X)$ has no roots in the field F.*

Taking into account Corollary 4.2.2 we obtain

Corollary 4.2.7. *Let F be a field, and let $f(X) \in F[X]$. Then the number of roots of the polynomial $f(X)$ belonging to the field F does not exceed* $\mathbf{deg}\, f(X)$.

Note that some of the roots $c_1, \dots, c_t$ may be the same. We say that the root c of the polynomial $f(X)$ is a multiple root of multiplicity m, if $f(X) = (X-c)^m g(X)$ where c is not a root of the polynomial $g(X)$. We can now refine Corollary 4.2.6 as follows.

Corollary 4.2.8. *Let F be a field, and let $f(X) \in F[X]$. Then $f(X) = q(X)$ $(X-c_1)^{r_1}\dots(X-c_t)^{r_t}$, where $c_j \neq c_k$ whenever $j \neq k$ and the polynomial $q(X)$ has no roots in the field F.*

A field F is called *algebraically closed* if every polynomial $f(X) \in F[X]$ has a root in F.

Corollary 4.2.9. *Let F be a field, and let $f(X) \in F[X]$. If F is algebraically closed, then $f(X) = (X-c_1)^{r_1}\dots(X-c_t)^{r_t}$ where $c_j \neq c_k$ whenever $j \neq k$.*

We state without proof the following important result proved by C. F. Gauss. This result has long been known as the Fundamental Theorem of Algebra.

Theorem 4.2.10. *The field $\mathbb{C}$ of complex numbers is algebraically closed.*

Certain polynomials play a very important role in polynomial rings. These are the irreducible polynomials. They are very much analogous to the prime numbers that are so well-known in the integers and play a similar role.

Definition 4.2.11. *Let R be an integral domain (in particular, a field). A polynomial $f(X) \in R[X]$ such that $\mathbf{deg}(f(X)) \geq 1$ is called irreducible or indecomposable over R, if it cannot be written as a product $f(X) = u(X)v(X)$, where $u(X), v(X) \in R[X]$ and $0 < \mathbf{deg}\, u(X), \mathbf{deg}\, v(X) < \mathbf{deg}\, f(X)$. A polynomial that is not irreducible is called reducible.*

This definition implies that every polynomial of degree 1 is irreducible. When discussing irreducible polynomials in some ring $R[X]$ the role of the ring R is important. For example, the polynomial $X^2 - 2$ is irreducible over the field $\mathbb{Q}$—it clearly does not factor into polynomials of smaller degree with coefficients in $\mathbb{Q}$—but over $\mathbb{R}$ we have the decomposition $X^2 - 2 = (X - \sqrt{2})(X + \sqrt{2})$. Thus, in $\mathbb{Q}[X]$, $X^2 - 2$ is irreducible, but in $\mathbb{R}[X]$ it is reducible. Likewise the polynomial $X^4 + 4$ can be factored over $\mathbb{Q}$ as

$$X^4 + 4 = (X^2 + 2X + 2)(X^2 - 2X + 2).$$

Both factors of this decomposition are irreducible not only over $\mathbb{Q}$ but also over $\mathbb{R}$. Finally, the polynomial $X^2 + 1$ is irreducible over $\mathbb{R}$, but

$$X^2 + 1 = (X + i)(X - i)$$

over the field of complex numbers.

The following theorems show the role of irreducible polynomials over fields. These theorems are analogous to the Fundamental Theorem of Arithmetic, that every natural number larger than 1 is a product of prime numbers uniquely.

Theorem 4.2.12. *Let F be a field. Every polynomial $f(X) \in F[X]$ such that $\mathbf{deg} f(X) \geq 1$ can be decomposed into a product of irreducible polynomials.*

Proof. We proceed by induction on $\mathbf{deg} f(X)$. If $\mathbf{deg} f(X) = 1$, then $f(X)$ is irreducible, and the result follows. Suppose now that $\mathbf{deg} f(X) > 1$, and suppose that we have proved our theorem for all polynomials $g(X) \in F[X]$ such that $\mathbf{deg}\, g(X) < \mathbf{deg}\, f(X)$. If $f(X)$ is irreducible, then again the theorem holds. If $f(X)$ is reducible, then $f(X) = g_1(X)g_2(X)$ where $g_1(X), g_2(X) \in F[X]$ and $1 < \mathbf{deg}\, g_1(X), \mathbf{deg}\, g_2(X) < \mathbf{deg}\, f(X)$. By our induction hypothesis $g_1(X) = h_1(X) \dots h_k(X)$ and $g_2(X) = h_{k+1}(X) \dots h_t(X)$, where $h_1(X), \dots, h_t(X) \in F[X]$ and $h_1(X), \dots, h_t(X)$ are irreducible polynomials over F. The result follows.

We supplement this theorem with the following uniqueness theorem, which is not needed in the sequel and whose proof is omitted.

Theorem 4.2.13. *Let F be a field and let $f(X)$ be an arbitrary polynomial over F such that* $\mathbf{deg}\, f(X) \geq 1$. *Suppose that*

$$f(X) = g_1(X) \dots g_k(X) = h_1(X) \dots h_t(X),$$

where $g_1(X), \dots, g_k(X), h_1(X), \dots, h_t(X) \in F[X]$ are irreducible over F. Then $k = t$ and there exists a permutation $\sigma \in \mathbf{S}_k$ such that $h_j(X) = u_{\sigma(j)} g_{\sigma(j)}(X)$ for some nonzero elements $u_1, \dots, u_k \in F$.

Theorem 4.2.13 can be revised as follows. A polynomial $f(X) = a_0 + a_1 X + \dots + a_n X^n$ over a ring R is called *unitary or monic* if $a_n = 1_R$.

Corollary 4.2.14. *Let F be a field and let $f(X)$ be an arbitrary polynomial over F such that* $\mathbf{deg}\,(f(X)) \geq 1$. *Then $f(X)$ can be written in the form*

$$f(X) = a(g_1(X))^{k_1} \dots (g_t(X))^{k_t},$$

where $g_1(X), \dots, g_k(X) \in F[X]$ and $g_1(X), \dots g_k(X)$ are irreducible monic polynomials over F, $g_j(X) \neq g_m(X)$ whenever $j \neq m$, and a is the leading coefficient of $f(X)$. Furthermore this representation is unique up to the order of the factors.

For polynomials over fields, Theorem 4.2.3 is analogous to Theorem 2.2.1 and, as we have seen, the ring of integers and the ring of polynomials over a field share many other similar properties concerned with divisibiltiy. We now consider another of these important properties.

Definition 4.2.15. *Let F be a field, and let $f(X), h(X) \in F[X]$. A polynomial $g(X)$ is called the greatest common divisor of $f(X)$ and $h(X)$ if it satisfies the following conditions:*

(GCD 1) *$g(X)$ divides both $f(X)$ and $h(X)$;*

(GCD 2) *if $q(X)$ is a common divisor of the polynomials $f(X), h(X)$, then $q(X)$ divides $g(X)$.*

An immediate question arises as to whether the great common divisor of every pair of polynomials exists, and then the question of uniqueness also arises.

The second question can be answered immediately: if $g(X)$ is a greatest common divisor of $f(X)$ and $h(X)$ and u is a nonzero element of F, then the polynomial $ug(X)$ satisfies the conditions (**GCD 1**), (**GCD 2**). Conversely, let $g_1(X)$ be a polynomial, satisfying the conditions (**GCD 1**), (**GCD 2**). Since $g_1(X)$ is a common divisor of $f(X)$ and $h(X)$, it follows that $g_1(X)$ divides $g(X)$

by (**GCD 2**). On the other hand, for the same reasons $g(X)$ divides $g_1(X)$. As we have seen above, this means that $g_1(X) = ug(X)$ for some nonzero element $u \in F$.

The question of the existence of the great common divisor for every pair $f(X), h(X)$ of polynomials, and the question of how to find this great common divisor, is solved by the Euclidean algorithm, just as was done in Section 2.2 for the integers. We now address this and note that the chain of reasoning is essentially the same as used for the Euclidean Algorithm used in the set of integers.

If $h(X) = 0_F$, then $\mathbf{GCD}(f(X), h(X)) = f(X)$. Therefore we can assume that $h(X) \neq 0_F$. Without loss of generality we can suppose that $\mathbf{deg}\, f(X) \geq \mathbf{deg}\, h(X)$. By Theorem 4.2.3, $f(X) = h(X)q_1(X) + r_1(X)$, where either $r_1(X) = 0_F$ or $\mathbf{deg}\, r_1(X) < \mathbf{deg}\, h(X)$. If $r_1(X) \neq 0_F$, then again using Theorem 4.2.3, we obtain $h(X) = r_1(X)q_2(X) + r_2(X)$, where either $r_2(X) = 0_F$ or $\mathbf{deg}\, r_2(X) < \mathbf{deg}\, r_1(X)$. If $r_2(X) \neq 0_F$, then divide $r_1(X)$ by $r_2(X)$ in the same way and continue this process until a remainder term is obtained that is 0_F. The process must terminate since at each step $\mathbf{deg}\, r_{j-1}(X) < \mathbf{deg}\, r_j(X)$. This means that the remainder at some appropriate stage, $r_k(X)$, should be zero. Thus, we obtain the following chain of equalities.

$$
\begin{aligned}
f(X) &= h(X)q_1(X) + r_1(X),\\
h(X) &= r_1(X)q_2(X) + r_2(X),\\
r_1(X) &= r_2(X)q_3(X) + r_3(X),\\
&\cdots,\\
r_{k-3}(X) &= r_{k-2}(X)q_{k-1}(X) + r_{k-1}(X),\\
r_{k-2}(X) &= r_{k-1}(X)q_k(X) + r_k(X),\\
r_{k-1}(X) &= r_k(X)q_{k+1}(X).
\end{aligned}
$$

We have

$$
\begin{aligned}
r_{k-2}(X) &= r_{k-1}(X)q_k(X) + r_k(X) = r_k(X)q_{k+1}(X)q_k(X) + r_k(X)\\
&= r_k(X)(q_{k+1}(X)q_k(X) + 1_F),
\end{aligned}
$$

so that $r_k(X)$ divides $r_{k-2}(X)$. Furthermore,

$$
\begin{aligned}
r_{k-3}(X) &= r_{k-2}(X)q_{k-1}(X) + r_{k-1}(X)\\
&= r_k(X)(q_{k+1}(X)q_k(X) + 1_F)q_{k-1}(X) + r_k(X)q_{k+1}(X)\\
&= r_k(X)(q_{k+1}(X)q_k(X)q_{k-1}(X) + q_{k-1}(X) + q_{k+1}(X)),
\end{aligned}
$$

so that $r_k(X)$ divides $r_{k-3}(X)$. Continuing further, moving up the chain of equalities, we eventually obtain that $r_k(X)$ divides $h(X)$ and $f(X)$. Let $v(X)$ be an arbitrary common divisor of $f(X)$ and $h(X)$. The equation $r_1(X) = f(X) - h(X)q_1(X)$ shows that $v(X)$ divides $r_1(X)$. The next equation $r_2(X) = h(X) - r_1(X)q_2(X)$ shows that $v(X)$ divides $r_2(X)$. Continuing further, moving down the chain, we obtain at the end, $v(X)$ divides $r_k(X)$. This means that $r_k(X)$ is the greatest common divisor of $f(X)$ and $h(X)$.

Exercise Set 4.2

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

4.2.1. What are the values of a, b, c such that the polynomials $f(X), g(X) \in \mathbb{Z}[X]$ are equal if $f(X) = aX^2(X+1) + b(X^2+1)(X-6) + cX(X^2+1)$, $g(X) = X^2 + 5X + 6$?

4.2.2. What are the values of a, b, c such that the polynomials $f(X), g(X) \in \mathbb{Z}[X]$ are equal if $f(X) = aX^2(X+3) + b(X-1)(X-6) + c(X+1)$, $g(X) = 2X^3 + 5X^2 + 8X - 5$?

4.2.3. For which value(s) of a is the polynomial $f(X) \in \mathbb{Z}[X]$, $f(X) = X^4 + 12X^3 + 38X^2 + aX + 1$ the square of some polynomial $g(X) \in \mathbb{Z}[X]$. Find all such $g(X)$.

4.2.4. For which values of a, b is the polynomial $f(X) \in \mathbb{Z}[X]$, $f(X) = X^3 + 6X^2 + aX + 8$ the cube of some polynomial $g(X) \in \mathbb{Z}[X]$.

4.2.5. Prove that the ring $\mathbb{Z}[X]$ contains no polynomial $f(X)$ such that $f(7) = 11$ and $f(11) = 13$.

4.2.6. For which values of a does the polynomial $g(X) = X^2 - a \in \mathbb{Z}[X]$ divide the polynomial $f(X) = 3X^4 - 2X^2 - 5 \in \mathbb{Z}[X]$?

4.2.7. For which values of a does the polynomial $g(X) = X^2 - a \in \mathbb{Q}[X]$ divide the polynomial $f(X) = 3X^4 - 2X^2 - 5 \in \mathbb{Q}[X]$?

4.2.8. For which values of a, b, c does the polynomial $g(X) = X^2 + a - 1 \in \mathbb{R}[X]$ divide the polynomial $f(X) = X^3 - bX - c \in \mathbb{R}[X]$?

4.2.9. Divide the polynomial $f(X) = 4X^4 - 6X^2 + 2X - 4$ by $g(X) = 2X^2 - 5X + 1$ in the ring $\mathbb{Q}[X]$.

4.2.10. Divide the polynomial $f(X) = 3X^3 - 4X + 24$ by $g(X) = X^2 + 1$ in the ring $\mathbb{R}[X]$.

4.2.11. Prove the rational root test which runs as follows: Let $f(X) = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$ be a polynomial with integer coefficients. Suppose that r is a rational number which is a root of this polynomial. Then r must be of the form a/b where a is a factor of a_0 and b is a factor of a_n.

4.2.12. Use the Euclidean algorithm to find $\mathbf{GCD}(f(X),g(X))$, if $f(X), g(X) \in \mathbb{Q}[X]$ where $f(X) = X^4 - X^3 - 2X^2 + 4X - 2, g(X) = X^3 - X^2 + X - 1$. Then find polynomials $u(X), v(X)$ such that

$$u(X)f(X) + v(X)g(X) = \mathbf{GCD}(f(X), g(X)).$$

4.2.13. Use the Euclidean algorithm to find $\mathbf{GCD}(f(X),g(X))$, if $f(X), g(X) \in \mathbb{Q}[X]$ where $f(X) = 3X^4 - 3X^3 + 4X^2 - X + 1, g(X) = 2X^3 - X^2 + X + 1$. Then find polynomials $u(X), v(X)$ such that

$$u(X)f(X) + v(X)g(X) = \mathbf{GCD}(f(X), g(X)).$$

4.2.14. Let $f(X), g(X) \in \mathbb{Q}[X]$ where $f(X) = X^5 + 2X^4 - 3X^3 + 6X - 4, g(X) = X^4 + 2X^3 - X^2 + 2X - 2$. Find the polynomials $u(X), v(X) \in \mathbb{Q}[X]$ such that

$$\mathbf{GCD}(f(X), g(X)) = u(X)f(X) + v(X)g(X).$$

4.2.15. Find the rational roots of the polynomial $X^4 + 2X^3 - 13X^2 - 38X - 24$.

4.2.16. Find the rational roots of the polynomial $2X^3 + 3X^2 + 6X - 4$.

4.2.17. Let $f(X) = X^5 + 4X^4 + 7X^3 + 8X^2 + 5X + 2 \in \mathbb{R}[X]$. Find all the irreducible factors of the polynomial $f(X)$ (over the ring $\mathbb{R}$), together with their multiplicity.

4.2.18. Prove that the polynomial equation $X^2 - 2X + 1 = 0$ has infinitely many solutions in $\mathbf{M}_2(\mathbb{R})$.

4.2.19. Prove that the polynomial $f(X) = X^3 - X^2 + X + 1$ is irreducible in the ring $\mathbb{Z}[X]$.

4.2.20. Prove that the polynomial $X^4 + 1$ is irreducible over $\mathbb{Q}$.

4.3 IDEALS AND QUOTIENT RINGS

The concept of a normal subgroup is one of the central ideas in group theory. Normal subgroups allow us to connect a group to some new related groups, the factor (or quotient) groups, and in turn these factor groups are related to

group homomorphisms. A similar situation arises in ring theory where the role of normal subgroups is played by special subsets called ideals.

Historically the concept of an ideal came from other needs and problems, and it was only later that the analogy between the concept of an ideal and a normal subgroup was recognized. In fact ideals were first studied in connection with rings in which uniqueness of factorization into products of prime elements did not hold. The search to discover some form of "weak uniqueness" property led mathematicians to the concept of an ideal. Ideals were first introduced by Dedekind in 1876 in the third edition of his book Vorlesungen über Zahlentheorie (in English: Lectures on Number Theory). The theory of ideals, a generalization of the concept of ideal numbers developed by Ernest Kummer, was later expanded by David Hilbert and especially Emmy Noether. For us there is no necessity to repeat all the steps in the development of the concept of an ideal. We give the definition of an ideal in its current-day form.

Definition 4.3.1. *A subring H of a ring R is called an ideal of R if, for each element $x \in R$ and every element $h \in H$, both products xh and hx lie in H. Write $H \lhd R$ to indicate that H is an ideal of R.*

We recall that, in this book, subrings do not have to contain the multiplicative identity of the ring. Indeed, as we show in Proposition 4.3.4 below, an ideal containing the identity element must be the whole ring. Thus proper ideals never contain the multiplicative identity of the ring.

The concept of an ideal is often developed using more general notions as follows. A subring H of a ring R is called a left ideal (respectively right ideal), if for each element $x \in R$ and every element $h \in H$ the product xh (respectively hx) lies in H. We write $H \lhd_l R$ to denote that H is a left ideal of R and $H \lhd_r R$ to denote that H is a right ideal of R. Every subset of the ring, which is both a left and a right ideal simultaneously, is an ideal. Thus one can talk about the "two-sided" ideals of a ring. If R is a commutative ring, then all these concepts coincide. In every ring R the subsets $\{0_R\}$ and R are always ideals, as is easily verified.

Using Theorem 4.1.4 we obtain the following criterion for a subset to be an ideal. This is often used as the working definition of ideal.

Proposition 4.3.2. *Let R be a ring. A nonempty subset H of a ring R is an ideal if and only if the following conditions hold:*

(I 1) *if $x, y \in H$, then $x - y \in H$;*

(I 2) *if $x \in R$ and $h \in H$ then the products xh and hx both belong to H.*

The following statement follows directly from Proposition 4.3.2. Its proof is similar to the proof of Corollary 4.1.5 and we omit it.

Corollary 4.3.3. *Let R be a ring and let $\mathfrak{S}$ be a family of ideals of R. The intersection $\cap\mathfrak{S}$ of all ideals from this family is also an ideal of R.*

In a commutative ring, R, the simplest ideals arise in the following way. Let $a \in R$ and let $aR = \{ax \mid x \in R\}$. Clearly $aR \neq \emptyset$. We show that aR satisfies the conditions **(I 1)**, **(I 2)**. Indeed, if $u, v \in aR$, then $u = ay, v = az$ for some elements $y, z \in R$. We have

$$u - v = ay - az = a(y - z) \in aR, \text{ and}$$
$$ru = ur = (ay)r = a(yr) \in aR$$

for an arbitrary element r of the ring R. Thus aR is an ideal of R. Since $a = a1_R$, then $a \in aR$. The ideal aR is called a principal ideal (or the principal ideal generated by a) of the commutative ring R.

In Section 4.1 we showed that if H is a subring of $\mathbb{Z}$, then $H = n\mathbb{Z}$ for some nonnegative integer n. Hence every subring of $\mathbb{Z}$ is a principal ideal.

Proposition 4.3.4. *Let R be a ring and let H an ideal of R. If H contains an invertible element of R, then $H = R$.*

Proof. Suppose that $x \in H \cap \mathbf{U}(R)$. Since H is an ideal, $1_R = x^{-1}x \in H$ and if a is an arbitrary element of R, then $a = a1_R \in H$. This means that $H = R$.

Definition 4.3.5. *A ring R is called simple if its only ideals are $\{0_R\}$ and R.*

Proposition 4.3.4 immediately implies the following results.

Corollary 4.3.6. *Let R be a division ring (in particular, let R be a field). If H is an ideal of R, then either $H = \{0_R\}$ or $H = R$.*

Corollary 4.3.7. *Every division ring, and hence every field, is a simple ring.*

Although Corollary 4.3.7 gives us many examples of simple rings, we show next that, for commutative rings, there are no other examples of simple rings.

Theorem 4.3.8. *Every simple commutative ring R is a field.*

Proof. Let a be an arbitrary nonzero element of R. The nonzero ideal aR contains a and hence $aR = R$, since R is simple. Therefore there is an element $x \in R$ such that $xa = ax = 1_R$, so a is invertible. Consequently every nonzero element is invertible, so that R is a field.

As we will see a little bit later, the situation is much more complicated for noncommutative rings. As the ring of Quaternions shows there are simple rings that are not fields.

We now define the sum of two subrings. Accordingly, let R be a ring, and let H, K be subrings of R. Put

$$H + K = \{x + y \mid x \in H, y \in K\}.$$

The subset $H + K$ is called *the sum* of the subrings H, K.

For elements $x \in H, y \in K$ we have $x = x + 0_R \in H + K, y = 0_R + y \in H + K$. This shows that $H, K \subseteq H + K$. We remark that a sum of subrings is not always a subring since products of elements in $H + K$ need not be in $H + K$. However if one or both of H, K is an ideal of R, then $H + K$ is a subring. Suppose, for example, that H is an ideal of R and let $a, b \in H + K$. Then $a = x + y, b = x_1 + y_1$ for certain elements $x, x_1 \in H, y, y_1 \in K$. We have

$$a - b = (x + y) - (x_1 + y_1) = (x - x_1) + (y - y_1) \in H + K,$$

and

$$ab = (x + y)(x_1 + y_1) = xx_1 + xy_1 + yx_1 + yy_1.$$

Since H is an ideal, $xx_1 + xy_1 + yx_1 \in H$. Also since K is a subring, $yy_1 \in K$ so that,

$$(xx_1 + xy_1 + yx_1) + yy_1 \in H + K.$$

Hence $H + K$ satisfies the conditions (**SR 1**) and (**SR 2**), and Theorem 4.1.4 shows that $H + K$ is a subring of R.

Furthermore, if H, K are ideals of R, then the sum $H + K$ is also an ideal. In this case in fact, if $x \in H, y \in K$, $z \in R$, then

$$z(x + y) = zx + zy \in H + K, (x + y)z = xz + yz \in H + K,$$

so that $H + K$ satisfies the conditions (**I 1**), (**I 2**), and by Proposition 4.3.2, it is an ideal of R. Note also that if an ideal L of R contains the ideals H, K, then $H + K \leq L$.

As we mentioned above, every ideal H of the ring $\mathbb{Z}$ has the form $H = n\mathbb{Z}$ for some nonnegative n. In particular, each ideal of $\mathbb{Z}$ is principal. The following idea is therefore very natural.

Definition 4.3.9. *An integral domain R is called a Principal Ideal Domain (sometimes called a PID for short) if every ideal of R is principal; thus if $I \lhd R$ then $I = aR$ for some $a \in I$.*

Using Theorem 4.2.3 we can easily obtain further examples of PID.

Proposition 4.3.10. *Let F be a field. Then the polynomial ring $F[X]$ is a PID.*

Proof. Corollary 4.2.2 shows that $F[X]$ is an integral domain. Let H be an ideal of $F[X]$. If $H = \{0_F\}$, then $H = 0_F F[X]$, so H is a principal ideal. Assume now that H is nonzero. Let $\mathcal{S} = \{\mathbf{deg}\, f(X) \mid 0_F \neq f(X) \in H\}$, a nonempty subset of $\mathbb{N}_0$, so $\mathcal{S}$ has a least element d. Choose a polynomial $h(X) \in H$ such that $\mathbf{deg}\, h(X) = d$. Since H is an ideal of $F[X]$, $h(X)f(X) \in H$ for each polynomial $f(X) \in F[X]$. Thus $h(X)F[X] \leq H$. Now let $g(X)$ be an arbitrary element of H. By Theorem 4.2.3, there exist polynomials $q(X), r(X) \in F[X]$ such that $g(X) = q(X)h(X) + r(X)$, where either $r(X) = 0_F$ or $\mathbf{deg}\, r(X) < \mathbf{deg}\, h(X)$.

We have $r(X) = g(X) - q(X)h(X)$. As we noted above, $q(X)h(X) \in H$, so that $r(X) \in H$. If we suppose now that $r(X) \neq 0_F$, then $\mathbf{deg}\, r(X) < d = \mathbf{deg}\, h(X)$, and we obtain a contradiction to the choice of d. Therefore $r(X) = 0_F$. Hence $g(X) = q(X)h(X)$, which proves the equality $H = h(X)F[X]$. Thus $F[X]$ is a PID as claimed.

Using the idea of a principal ideal we can give the following characterization of the divisibility of polynomials over fields.

Proposition 4.3.11. *Let F be a field and let $f(X), g(X) \in F[X]$. Then $f(X)$ is a divisor of $g(X)$ if and only if $g(X)F[X] \leq f(X)F[X]$.*

Proof. Suppose that $f(X)$ is a divisor of $g(X)$. Then $g(X) = f(X)h(X)$ for some polynomial $h(X) \in F[X]$. Thus $g(X)q(X) = f(X)h(X)q(X) \in f(X)F[X]$ for each arbitrary polynomial $q(X) \in F[X]$. Thus $g(X)F[X] \leq f(X)F[X]$.

Conversely, suppose that $g(X)F[X] \leq f(X)F[X]$. Since $g(X) \in g(X)F[X]$ we deduce that $g(X) \in f(X)F[X]$, so $g(X) = f(X)h(X)$ for some polynomial $h(X) \in F[X]$. The result follows.

In Section 4.1 we considered the ring

$$\mathbb{Z}[\sqrt{r}] = \{a + b\sqrt{r} \mid a, b \in \mathbb{Z}\}$$

where r is an integer with the property $\sqrt{r} \notin \mathbb{Q}$. Setting $r = -1$, we obtain the ring $\mathbb{Z}[i]$, which is called the *ring of Gaussian integers*, named after C. F. Gauss, who first studied the arithmetic properties of this ring.

If $\alpha = a+bi$, where $a,b \in \mathbb{Q}$, then define the *norm* of α by $N(\alpha) = a^2+b^2 = |\alpha|^2$. This is simply the square of the modulus of α. Then

$$N(\alpha\beta) = |\alpha\beta|^2 = (|\alpha|\,|\beta|)^2 = |\alpha|^2\,|\beta|^2 = N(\alpha)N(\beta).$$

Proposition 4.3.12. *The ring $\mathbb{Z}[i]$ is a PID.*

Proof. Let $\alpha = a+bi$ and $\beta = c+di \neq 0$ where $a,b,c,d \in \mathbb{Z}$. Then $\frac{\alpha}{\beta} = x+yi$, where $x,y \in \mathbb{Q}$ and hence there exist integers u,v such that $|x-u| \leq \frac{1}{2}$ and $|y-v| \leq \frac{1}{2}$. Let $\gamma = u+vi$ and $\rho = \alpha - \beta\gamma$. By definition, $\gamma, \rho \in \mathbb{Z}[i]$ and either $\rho = 0$ or

$$\begin{aligned} N(\rho) = N(\alpha-\beta\gamma) &= N\left(\beta\left(\frac{\alpha}{\beta}-\gamma\right)\right) = N(\beta)N\left(\beta\left(\frac{\alpha}{\beta}-\gamma\right)\right) \\ &= N(\beta)N((x-u)+(y-v)i) = N(\beta)((x-u)^2+(y-v)^2) \\ &\leq N(\beta)\left(\frac{1}{4}+\frac{1}{4}\right) = \frac{1}{2}N(\beta) < N(\beta). \end{aligned}$$

Hence for each pair of elements $\alpha, \beta \in \mathbb{Z}[i]$, where $\beta \neq 0$, there are elements $\gamma, \rho \in \mathbb{Z}[i]$ such that $\alpha = \beta\gamma + \rho$ and $N(\rho) < N(\beta)$. Now let H be an ideal of $\mathbb{Z}[i]$. If $H = \{0\}$, then H is a principal ideal, so we may assume that H is nonzero. Let

$$\mathcal{S} = \{N(\beta) \mid 0 \neq \beta \in H\}.$$

Then $\mathcal{S}$ is a nonempty subset of $\mathbb{N}$, so has a least element d, say. Choose an element $\gamma \in H$ such that $N(\gamma) = d$. Since H is an ideal of $\mathbb{Z}[i]$, it follows that $\gamma\mathbb{Z}[i] \leq H$. Let α be an arbitrary element of H. From our discussion above, there exist elements $\delta, \rho \in \mathbb{Z}[i]$ such that $\alpha = \gamma\delta + \rho$ and $0 \leq N(\rho) < N(\gamma) = d$. We have $\rho = \alpha - \gamma\delta \in H$ and since $N(\rho) < d$ then $\rho = 0$, by the choice of γ. Hence $\alpha = \gamma\delta$, which proves the equality $H = \gamma\mathbb{Z}[i]$. Hence $\mathbb{Z}[i]$ is a PID.

Next, let p be a prime, and consider the subset $\mathbb{Q}_p = \left\{\frac{m}{p^k} \mid m,k \in \mathbb{Z}\right\}$ of $\mathbb{Q}$. We saw in Section 4.1 that $\mathbb{Q}_p$ is a subring of $\mathbb{Q}$ and we now describe the ideals of this ring. Let H be an ideal of $\mathbb{Q}_p$. If $H = \{0\}$, then H is a principal ideal. Assume now that H is nonzero. If $0 \neq \frac{m}{p^k} \in H$, then $0 \neq m = p^k\left(\frac{m}{p^k}\right) \in H$, so $H \cap \mathbb{Z} \neq \{0\}$. Let

$$\mathcal{S} = \{k \mid 0 < k \in H \cap \mathbb{Z}\}.$$

Then $\mathcal{S}$ is a nonempty subset of $\mathbb{N}$, so $\mathcal{S}$ has a least element d. Since H is an ideal of $\mathbb{Q}_p$, it follows that $d\mathbb{Q}_p \leq H$. On the other hand, let $\frac{n}{p^k}$ be an arbitrary element of H. We have $n = qd + r$ where $0 \leq r < d$. Then

$$\frac{n}{p^k} = \frac{(qd+r)}{p^k} = \frac{qd}{p^k} + \frac{r}{p^k} = \left(\frac{q}{p^k}\right) d + \frac{r}{p^k}.$$

As we noted above, $\left(\frac{q}{p^k}\right) d \in H$, so that $\frac{r}{p^k} = \frac{n}{p^k} - \frac{qd}{p^k} \in H$. Since H is an ideal, $r = p^k \left(\frac{r}{p^k}\right) \in H$. The choice of d and the inequality $r < d$ show that $r = 0$ and hence $\frac{n}{p^k} = \left(\frac{q}{p^k}\right) d$. This proves that $H = d\mathbb{Q}_p$ so $\mathbb{Q}_p$ is also a PID.

We now give some examples of ideals in noncommutative rings. First of all we consider the ring of matrices $\mathbf{M}_n(\mathbb{R})$. Let L be a nonzero ideal of the ring $\mathbf{M}_n(\mathbb{R})$ and suppose that $B = [b_{ij}]$ is a nonzero matrix which is an element of L. Then there are indices k, m such that $b_{km} \neq 0$. We write $B = \sum_{1 \leq i,j \leq n} b_{ij} E_{ij}$. Since $E_{st} E_{ij} = E_{sj}$, provided $t = i$, and is the zero matrix otherwise, it follows that

$$E_{st} B = E_{st} \left(\sum_{1 \leq i,j \leq n} b_{ij} E_{ij} \right) = \sum_{1 \leq i,j \leq n} b_{ij} E_{st} E_{ij} = \sum_{1 \leq j \leq n} b_{tj} E_{sj}.$$

$$\text{Hence } E_{st} B E_{qr} = \left(\sum_{1 \leq j \leq n} b_{tj} E_{sj} \right) E_{qr} = \sum_{1 \leq j \leq n} b_{tj} E_{sj} E_{qr} = b_{tq} E_{sr}.$$

Since L is an ideal of $\mathbf{M}_n(\mathbb{R})$, it follows that $E_{st} B E_{qr} \in L$ for any s, t, q, r. Now let $A = [a_{ij}]$ be an arbitrary matrix. Since $E_{ik} B E_{mj} = b_{km} E_{ij} \in L$ we see that

$$a_{ij} E_{ij} = (a_{ij} b_{km}^{-1} I)(E_{ik} B E_{mj}) \in L$$

for every pair of indices i, j and this proves that $A = \sum_{1 \leq i,j \leq n} a_{ij} E_{ij} \in L$. We deduce that $L = \mathbf{M}_n(\mathbb{R})$.

As we saw in Section 4.1, $\mathbf{M}_n(\mathbb{R})$ has zero-divisors, so it cannot be a division ring. Hence we have an example of a simple noncommutative ring that is not a division ring. Notice that this is in contrast to Theorem 4.3.8.

However the matrix ring $\mathbf{M}_n(\mathbb{R})$ contains nonsimple subrings. One of these is the subring $\mathbf{T}_n(\mathbb{R})$ of upper triangular matrices. Its subring $\mathbf{NT}_n(\mathbb{R})$, of zero-triangular matrices, is an ideal of $\mathbf{T}_n(\mathbb{R})$. Indeed, since $\mathbf{NT}_n(\mathbb{R})$ is a subring, it satisfies the condition **(I 1)**. Let $A = [a_{ij}] \in \mathbf{T}_n(\mathbb{R})$, $B = [b_{ij}] \in \mathbf{NT}_n(\mathbb{R})$ and let $AB = [c_{ij}]$. In Section 4.1 we proved that $AB \in \mathbf{T}_n(\mathbb{R})$ and that $c_{jj} = a_{jj} b_{jj}$,

for $1 \leq j \leq n$. Since $b_{jj} = 0$ for all j, it follows that $c_{jj} = 0$ for all j, and hence $AB \in \mathbf{NT}_n(\mathbb{R})$. Likewise $BA \in \mathbf{NT}_n(\mathbb{R})$. This shows that $\mathbf{NT}_n(\mathbb{R})$ satisfies the condition **(I 2)** and, using Proposition 4.3.2, we deduce that it is an ideal of $\mathbf{T}_n(\mathbb{R})$.

We note that neither the subring $\mathbf{D}_n(\mathbb{R})$ of all diagonal matrices, nor the subring $\mathbb{R}I = \{\lambda I \mid \lambda \in \mathbb{R}\}$ of all scalar matrices are ideals of $\mathbf{M}_n(\mathbb{R})$. For instance, the equality

$$\begin{pmatrix} a & b & c \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x & 0 & 0 \\ 0 & y & 0 \\ 0 & 0 & z \end{pmatrix} = \begin{pmatrix} ax & by & cz \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

shows that neither $\mathbf{D}_n(\mathbb{R})$ nor $\mathbb{R}I$ are ideals of $\mathbf{M}_3(\mathbb{R})$.

In linear algebra courses it may be shown that the scalar matrices and only the scalar matrices commute with all other matrices. In other words, the center of the ring $\mathbf{M}_n(\mathbb{R})$ coincides with $\mathbb{R}I$. This shows that the center of a ring is not necessarily an ideal of the ring.

As we mentioned above, the concept of an ideal is a ring's analogue of a normal subgroup. It is precisely when a subgroup H in a group G is normal that the set of all cosets of G by H is a group. We now consider the analogue of this construction in ring theory. Let R be a ring and let H be an ideal of R. By the definition of an ideal it follows that the additive group of H is a subgroup of the (abelian) additive group of R. Therefore we may construct the group of cosets of R modulo H, since H will be a normal subgroup of R under the operation of addition. Since we are thinking of R as a group under addition at the moment the cosets of R/H are denoted by $x+H$ (or $H+x$) and the operation of addition in R/H is defined by $(x+H)+(y+H) = x+y+H$ for all $x, y \in R$.

As we saw in Section 3.3 R/H becomes an abelian group under this addition. To transform R/H into a ring, we define the rule for multiplying cosets by $(x+H)(y+H) = xy+H$ for all $x, y \in R$.

First we must show that this rule is well-defined, namely that it is independent of the choice of representatives of the cosets $x+H$ and $y+H$. Indeed, let $x_1+H = x+H$ and $y_1+H = y+H$, where $x, y, x_1, y_1 \in R$. Then $x_1 = x+h_1, y_1 = y+h_2$ for certain elements $h_1, h_2 \in H$. We have

$$x_1y_1 = (x+h_1)(y+h_2) = xy+xh_2+h_1y+h_1h_2.$$

Since H is an ideal, $xh_2, h_1y, h_1h_2 \in H$, so that $x_1y_1 \in xy+H$, which means that $x_1y_1+H = xy+H$.

Furthermore, for all $x, y, z \in R$ we have

$$\begin{aligned}(x+H)((y+H)+(z+H)) &= (x+H)(y+z+H) = x(y+z)+H \\ &= xy+xz+H = (xy+H)+(xz+H) \\ &= (x+H)(y+H)+(x+H)(z+H).\end{aligned}$$

Similarly we can prove the equation

$$((x+H)+(y+H))(z+H) = (x+H)(z+H)+(y+H)(z+H),$$

and the equation

$$\begin{aligned}(x+H)((y+H)(z+H)) &= (x+H)(yz+H) = x(yz)+H \\ &= (xy)z+H = (xy+H)(z+H) \\ &= ((x+H)(y+H))(z+H).\end{aligned}$$

These computations show that the distributive and associative properties hold in R/H. Also

$$(1_R+H)(x+H) = 1_R x+H = x+H, (x+H)(1_R+H) = x1_R+H = x+H.$$

These equations show that R/H is a ring with these operations and the coset 1_R+H is its identity element.

Definition 4.3.13. *Let R be a ring and let H be an ideal of R. The set of all cosets of the ideal H, using the operations defined above, is called the quotient (or factor) ring of R over H and is denoted by R/H.*

If R is a commutative ring then, for all $x, y \in R$, we have

$$(x+H)(y+H) = xy+H = yx+H = (y+H)(x+H),$$

so that R/H is also commutative.

As we mentioned earlier, every nonzero ideal of the ring $\mathbb{Z}$ has the form $n\mathbb{Z}$ where $n \neq 0$. In Section 3.2 we showed that $|\mathbb{Z}/n\mathbb{Z}| = n$. We observe the following important properties of the quotient ring $\mathbb{Z}/n\mathbb{Z}$.

Proposition 4.3.14. *If n is a prime, then $\mathbb{Z}/n\mathbb{Z}$ is a field. If n is not prime, then $\mathbb{Z}/n\mathbb{Z}$ has zero-divisors.*

Proof. Let n be a prime and suppose that $x+n\mathbb{Z} \neq n\mathbb{Z}$ where $x \in \mathbb{Z}$. Then $\mathbf{GCD}(x,n) = 1$ and, by Corollary 2.2.7, there exist integers u, v such that $xu + nv = 1$. We now have

$$\begin{aligned} 1+n\mathbb{Z} = xu+nv+n\mathbb{Z} &= (xu+n\mathbb{Z})+(nv+n\mathbb{Z}) \\ &= (x+n\mathbb{Z})(u+n\mathbb{Z})+(0+n\mathbb{Z}) = (x+n\mathbb{Z})(u+n\mathbb{Z}). \end{aligned}$$

This shows that every nonzero element of $\mathbb{Z}/n\mathbb{Z}$ is invertible, so $\mathbb{Z}/n\mathbb{Z}$ is a field.

Suppose now that n is not a prime. Then $n = km$ where $1 < k < n, 1 < m < n$. It follows that $k+n\mathbb{Z} \neq 0+n\mathbb{Z}$, and $m+n\mathbb{Z} \neq 0+n\mathbb{Z}$. At the same time, however,

$$(k+n\mathbb{Z})(m+n\mathbb{Z}) = km+n\mathbb{Z} = n+n\mathbb{Z} = 0+n\mathbb{Z}.$$

This proves that $k+n\mathbb{Z}$ and $m+n\mathbb{Z}$ are zero-divisors, as required.

Primes in the integers motivate a further definition, which we now discuss. Let $p \in \mathbb{Z}$ be a prime and let H be an ideal of $\mathbb{Z}$ with the properties $p\mathbb{Z} \leq H$ and $p\mathbb{Z} \neq H$. Then $H \setminus p\mathbb{Z} \neq \emptyset$. Choose an integer $k \in H \setminus p\mathbb{Z}$. Since $k \notin p\mathbb{Z}$, p is not a divisor of k. It follows that $\mathbf{GCD}(k,p) = 1$. By Corollary 2.2.7, there exist integers u, v such that $ku + pv = 1$. Since H is an ideal, $ku \in H$. The inclusion $p\mathbb{Z} \leq H$ implies that $pv \in H$, so that $1 \in H$. But in this case, $H = \mathbb{Z}$, by Proposition 4.3.4. Hence there is no proper ideal strictly between $p\mathbb{Z}$ and $\mathbb{Z}$. This prompts the following definition.

Definition 4.3.15. *Let R be a ring. An ideal M of R is called maximal, if $M \neq R$ and whenever H is an ideal containing M either $H = M$ or $H = R$.*

As we saw earlier if p is a prime, then the ideal $p\mathbb{Z}$ is a maximal ideal of $\mathbb{Z}$. Maximal ideals in commutative rings are one means of constructing a field, in the following way.

Proposition 4.3.16. *Let R be a commutative ring. If M is a maximal ideal of R, then the quotient ring R/M is a field.*

Proof. Of course R/M is a commutative ring and $1_R + M$ is the multiplicative identity of this quotient ring. It suffices to prove that each nonzero element of R/M has a multiplicative inverse. Let x be an element of R such that $x+M \neq M$. Then $x \notin M$. As we saw earlier, the sum $xR+M$ is an ideal of R. Moreover, $x \notin M$ so that $xR+M \neq M$. Since M is a maximal ideal, $xR+M = R$. Then $1_R \in xR+M$, so $1_R = xy+z$ for certain elements $y \in R, z \in M$. Now we have

$$1_R + M = xy+z+M = xy+M = (x+M)(y+M).$$

This equation shows that every nonzero element of R/M is invertible, so that R/M is a field.

The next assertion shows the connection between the maximal ideals and irreducible polynomials in the ring of polynomials.

Proposition 4.3.17. *Let F be a field and let M an ideal of $F[X]$. Then M is maximal in $F[X]$ if and only if $M = p(X)F[X]$ where $p(X)$ is an irreducible polynomial over F.*

Proof. Suppose that M is a maximal ideal of $F[X]$. By Proposition 4.3.10 there exists a polynomial $p(X) \in F[X]$ such that $M = p(X)F[X]$. We remark that $\mathbf{deg}\, p(X) > 0$, otherwise $p(X) \in F$ and $M = F[X]$, contrary to M being maximal.

Assume that $p(X)$ is not irreducible over F. Then $p(X) = g(X)h(X)$ where $g(X), h(X) \in F[X]$ and $0 < \mathbf{deg}\, g(X), \mathbf{deg}\, h(X) < \mathbf{deg}\, p(X)$. We have $p(X)F[X] \leq g(X)F[X]$. Since $\mathbf{deg}\, g(X) < \mathbf{deg}\, p(X)$, it follows that $g(X)F[X] \nleq p(X)F[X]$. We deduce that $p(X)F[X] \neq g(X)F[X]$, and the maximality of M implies that $g(X)F[X] = F[X]$. Since $1_F \in F[X]$ it follows that $1_F = g(X)k(X)$ for some $k(X) \in F[X]$ and hence $\mathbf{deg}\, g(X) = 0$. This is a contradiction which shows that $p(X)$ must be irreducible.

Conversely, suppose that $p(X)$ is irreducible over F and let H be an ideal of $F[X]$ such that $p(X)F[X] \leq H \leq F[X]$. By Proposition 4.3.10, there exists a polynomial $h(X) \in F[X]$ such that $H = h(X)F[X]$. The inclusion $p(X)F[X] \leq h(X)F[X]$ implies that $p(X) = h(X)q(X)$ for some $q(X) \in F[X]$. Since $p(X)$ is irreducible, then either $\mathbf{deg}\, h(X) = 0$ or $\mathbf{deg}\, q(X) = 0$. If $\mathbf{deg}\, h(X) = 0$, then $h(X)F[X] = F[X]$; if $\mathbf{deg}\, q(X) = 0$, then $h(X)F[X] = p(X)F[X]$. This shows that H is a maximal ideal of $F[X]$. The result follows.

The notion of a quotient ring is closely connected to the concept of a ring homomorphism, which we discuss in the next section and where further examples of quotient rings are given. Indeed as an informal example, if n is a natural number, the ring $\mathbb{Z}/n\mathbb{Z} = \{0+n\mathbb{Z}, 1+n\mathbb{Z}, 2+n\mathbb{Z}, \ldots, (n-1)+n\mathbb{Z}\}$ is often identified with the set $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$ where multiplication and addition are defined modulo n and we shall use this alternative notation in future.

Exercise Set 4.3

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

4.3.1. On the set $R = \mathbb{Q} \times \mathbb{Z}$ define the operations of addition and multiplication by $(a,b)+(a_1,b_1) = (a+a_1, b+b_1), (a,b)(a_1,b_1) = (aa_1, bb_1)$. Prove that R is a ring with identity. Find all the ideals of R.

4.3.2. On the set $R = \mathbb{R} \times \mathbb{Z}$ we define the operations of addition and multiplication by $(a,b)+(a_1,b_1) = (a+a_1, b+b_1), (a,b)(a_1,b_1) = (aa_1, bb_1)$. Prove that R is a ring with identity. Find all the ideals of R.

4.3.3. Let R be a commutative ring, and let $n \in \mathbb{N}$. Is the set $nR = \{nx \mid x \in R\}$ an ideal of R?

4.3.4. Write the multiplication table for the ring $\mathbb{Z}_8 = \mathbb{Z}/8\mathbb{Z}$.

4.3.5. Write the multiplication table for the ring $\mathbb{Z}_5 = \mathbb{Z}/5\mathbb{Z}$.

4.3.6. Let $M = \{2k+2ti \mid k,t \in \mathbb{Z}\}$. Prove that M is an ideal of the ring $\mathbb{Z}[i]$. Find all elements of the factor ring $\mathbb{Z}[i]/M$.

4.3.7. Consider the ring $R = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mid a,b,c \in \mathbb{Z} \right\}$. Prove that the set $I = \left\{ \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \mid b \in \mathbb{Z} \right\}$ is an ideal of R.

4.3.8. Prove that $R = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mid a,b,c \in \mathbb{Z} \right\}$ is not an ideal of $\mathbf{M}_2(\mathbb{Z})$.

4.3.9. Let $M = \{2k+2ti \mid k,t \in \mathbb{Z}\}$. Find all zero-divisors of the quotient ring $\mathbb{Z}[i]/M$.

4.3.10. Prove that every subring of $\mathbb{Z}$ is an ideal of $\mathbb{Z}$. Hence describe the ideals of $\mathbb{Z}$.

4.3.10. Prove that in $\mathbf{M}_2(\mathbb{R})$ the set of matrices

$$\left\{ \begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix} | a,b \in \mathbb{R} \right\}$$

is a left ideal, but not a right ideal. Deduce that the left ideals of $\mathbf{M}_2(R)$ need not in general be of the form $\mathbf{M}_2(I)$, where R is a ring and I is an ideal of R.

4.3.12. Let R be a ring (with identity). Prove that the ideals of $\mathbf{M}_2(R)$ are of the form $\mathbf{M}_2(I)$, where I is an ideal of R.

4.3.13. Let R,S be rings (with identity). We can define a new ring $R \oplus S = \{(r,s) | r \in R, s \in S\}$ with operations defined by $(a,b)+(r,s) = (a+r, b+s)$ and $(a,b)(r,s) = (ar,bs)$ for all $a,r \in R, b,s \in S$. Let I be an ideal of R and let J be an ideal of S. Prove that $I \oplus J = \{(r,s) | r \in I, s \in J\}$ is an

ideal of $R \oplus S$. Then prove that if K is an ideal of $R \oplus S$ then $K = I \oplus J$ for some ideals I, J of R, S respectively.

4.3.14. Find all the irreducible polynomials of degree at most 4 in the ring $\mathbb{Z}_2[X]$.

4.3.15. Let p be a prime and let $\mathbf{GL}_2(\mathbb{Z}/p\mathbb{Z})$ denote the group of all invertible matrices with coefficients in $\mathbb{Z}/p\mathbb{Z}$. Prove that the natural map $f : \mathbf{GL}_2(\mathbb{Z}) \longrightarrow \mathbf{GL}_2(\mathbb{Z}/p\mathbb{Z})$ is a group homomorphism and find its kernel.

4.3.16. Find the greatest common divisor of the polynomials $X^4 + X^3 + X^2 + 2X + 1$ and $2X^3 + 3X^2 + 2X + 1$ in the ring $\mathbb{Z}_5[X]$. Then write this greatest common divisor as a linear combination of the two given polynomials, using the Euclidean Algorithm.

4.3.17. Find all the rings with unity in which every subring is an ideal.

4.3.18. Let R be a ring and let I be an ideal of R. Let $J = \{r \in R | ra = 0 \text{ for all } a \in I\}$. Prove that J is an ideal of R.

4.3.19. An element a of a ring R is called nilpotent if $a^n = 0$ for some natural number n. If R is a commutative ring then prove that the set $S = \{a \in R | a \text{ is nilpotent } \}$ is an ideal of R. Prove further that the only nilpotent element of R/S is $0 + S$. (S is called the nil radical of the ring R.)

4.3.20. Find the nil radical of the ring $\mathbb{Z}_{28}$.

4.4 HOMOMORPHISMS OF RINGS

Let R be a ring and let H be an ideal of R. Let $\sigma_H : R \longrightarrow R/H$ be the mapping defined by $\sigma_H(x) = x + H$ for every $x \in R$.

This mapping has the following important properties, obtained using the definitions of addition and multiplication in the ring R/H namely,

$$\sigma_H(x+y) = x+y+H = (x+H)+(y+H) = \sigma_H(x)+\sigma_H(y) \text{ and}$$
$$\sigma_H(xy) = xy+H = (x+H)(y+H) = \sigma_H(x)\sigma_H(y).$$

We shall say that the mapping respects both ring operations. This is a very natural concept and in this section we will study the basic properties of such mappings.

Definition 4.4.1. *Let R, S be rings. The mapping $f : R \longrightarrow S$ is called a ring homomorphism if it satisfies the conditions*

$$f(x+y)=f(x)+f(y) \text{ and } f(xy)=f(x)f(y)$$

for all elements $x, y \in R$.

In this section we only consider rings, so the term "homomorphism" will always be understood to mean "ring homomorphism." As usual an injective homomorphism is called a monomorphism, a surjective homomorphism is called an epimorphism, and a bijective homomorphism is called an isomorphism.

Suppose that $f : R \longrightarrow S$ is an isomorphism. Then f is bijective, so Corollary 1.3.5 shows that there exists an inverse mapping $f^{-1} : S \longrightarrow R$. Let $u, v \in S$, and $x = f^{-1}(u), y = f^{-1}(v)$. Then $u = f(x), v = f(y)$ and we have

$$\begin{aligned} f^{-1}(u+v) &= f^{-1}(f(x)+f(y)) = f^{-1}(f(x+y)) = x+y \\ &= f^{-1}(u)+f^{-1}(v), \text{ and} \\ f^{-1}(uv) &= f^{-1}(f(x)f(y)) = f^{-1}(f(xy)) = xy = f^{-1}(u)f^{-1}(v). \end{aligned}$$

This shows that f^{-1} is a homomorphism and also an isomorphism, since it is bijective.

Definition 4.4.2. *Let* R, S *be rings. Then* R *and* S *are called isomorphic if there exists an isomorphism from* R *to* S *or, equivalently, from* S *to* R. *This will be denoted by* $R \cong S$.

The easiest example of an isomorphism is the identity permutation $\varepsilon_R : R \longrightarrow R$, defined by $\varepsilon(r) = r$, for all $r \in R$. It is also easy to show that if $f : R \longrightarrow S$ and $g : S \longrightarrow U$ are homomorphisms, then their product $g \circ f$ is likewise a homomorphism.

The following results on ring homomorphisms are entirely analogous to those we obtained for groups.

Proposition 4.4.3. *Let* R, S *be rings and let* $f : R \longrightarrow S$ *be a homomorphism. Then the following properties hold:*

(i) $f(0_R) = 0_S$;
(ii) $f(-x) = -f(x)$ *for every element* $x \in R$;
(iii) $f(x-y) = f(x) - f(y)$ *for all* $x, y \in R$;
(iv) *If* H *is a subring of* R, *then its image,* $f(H)$, *is a subring of* S. *In particular,* $f(R) = \mathbf{Im} f$ *is a subring of* S;
(v) *If* V *is a subring of* S, *then its preimage* $f^{-1}(V)$ *is a subring of* R;

(vi) *If V is an ideal of S, then its preimage $f^{-1}(V)$ is an ideal of R. In particular,*

$$\mathbf{Ker} f = \{x \in R \mid f(x) = 0_S\} = f^{-1}(0_S)$$

is an ideal of R;

(vii) *If R has the multiplicative identity, 1_R, then $f(1_R)$ is the identity element of the subring* $\mathbf{Im} f$*;*

(viii) *If R is commutative, then* $\mathbf{Im} f$ *is commutative.*

(ix) *If f is an epimorphism and H is an ideal of R then $f(H)$ is an ideal of S.*

Proof.

(i) We have $x + 0_R = x$ for each $x \in R$. Then

$$f(x) + f(0_R) = f(x + 0_R) = f(x).$$

Since the element $f(x)$ has a negative in S, we may add this negative to both sides of these equations to obtain

$$0_S = -f(x) + f(x) = -f(x) + f(x) + f(0_R) = 0_S + f(0_R) = f(0_R).$$

(ii) From the definition of the negative element we have $x + (-x) = 0_R$, so that

$$0_S = f(0_R) = f(x + (-x)) = f(x) + f(-x).$$

This equation show that $f(-x)$ is the negative of $f(x)$.

(iii) We have

$$f(x - y) = f\big(x + (-y)\big) = f(x) + f(-y) = f(x) + \big(-f(y)\big) = f(x) - f(y).$$

(iv) Let $x, y \in H$ and let $a = f(x), b = f(y)$. Then

$$a - b = f(x) - f(y) = f(x - y) \in f(H) \text{ and}$$
$$ab = f(x)f(y) = f(xy) \in f(H).$$

It follows from Theorem 4.1.4 that the nonempty set $f(H)$ is a subring of S.

(v) Let $x, y \in f^{-1}(V)$. Then $f(x), f(y) \in V$. Since V is a subring of S, $f(x) - f(y) = f(x - y) \in V$, and $f(x)f(y) = f(xy) \in V$, which implies that

$x-y, xy \in f^{-1}(V)$. From Theorem 4.1.4 it follows that the nonempty subset $f^{-1}(V)$ is a subring of R.

(vi) As in the proof of (v), if $x, y \in f^{-1}(V)$ then $x-y \in f^{-1}(V)$. If also $r \in R$ then $f(xr) = f(x)f(r) \in V$, since V is an ideal of S, and hence $xr \in f^{-1}(V)$. Likewise $rx \in f^{-1}(V)$. Thus conditions **(I 1)** and **(I 2)** of Proposition 4.3.2 are satisfied and it follows that the nonempty set $f^{-1}(V)$ is an ideal of R. Of course $\{0_S\}$ is an ideal of S so the statement concerning $\mathbf{Ker}f$ is clear.

(vii) If $a \in \mathbf{Im}f$, then $a = f(x)$ for some element $x \in R$. Therefore

$$a = f(x) = f(x1_R) = f(x)f(1_R) = af(1_R) \text{ and}$$
$$a = f(x) = f(1_R x) = f(1_R)f(x) = f(1_R)a,$$

which means that $f(1_R)$ is the identity element of the subring $\mathbf{Im}f$.

(viii) If R is commutative, let $a, b \in \mathbf{Im}f$. Then $a = f(x)$ and $b = f(y)$ for some elements $x, y \in R$. We now have

$$ab = f(x)f(y) = f(xy) = f(yx) = f(y)f(x) = ba.$$

Hence $\mathbf{Im}f$ is commutative also.

(ix) By (iv) $f(H)$ is a subring of S. Let $x \in H$, let $a = f(x)$ and let $u \in S$. Since f is an epimorphism, there exists an element $y \in R$ such that $u = f(y)$. Then

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

Since H is an ideal of R, we have $xy \in H$ and hence $au \in f(H)$. Similarly we can prove that $ua \in f(H)$. Hence $f(H)$ satisfies the conditions of Definition 4.3.1 and therefore $f(H)$ is an ideal of S.

The ideal $\mathbf{Ker}(f)$ is called the kernel of the homomorphism f. Now we are in a position to prove some theorems concerning homomorphisms of rings that are analogous to the corresponding theorems about groups.

Theorem 4.4.4. *(The theorem on monomorphisms). Let R, S be rings. A homomorphism $f : R \longrightarrow S$ is a monomorphism if and only if $\mathbf{Ker}f = \{0_R\}$. If $f : R \longrightarrow S$ is a monomorphism, then $R \cong \mathbf{Im}f$.*

Proof. If f is a monomorphism, and if $x \neq 0_R$ then $f(x) \neq f(0_R) = 0_S$. This means that no nonzero element x belongs to $\mathbf{Ker}f$ and hence $\mathbf{Ker}f = \{0_R\}$.

Conversely, let $\mathbf{Ker}f = \{0_R\}$ and let x, y be elements of R such that $f(x) = f(y)$. Then

$$f(x-y) = f(x) - f(y) = 0_S$$

and hence $x - y \in \mathbf{Ker}f$. It follows that $x - y = 0_R$, so that $x = y$. Thus f is an injective homomorphism and hence it is a monomorphism.

Theorem 4.4.5. *(The First Isomorphism Theorem, version 1). Let R, S be rings and let $f : R \longrightarrow S$ be an epimorphism. Then S is isomorphic to $R/\mathbf{Ker}f$.*

Proof. For the sake of convenience put $H = \mathbf{Ker}f$. We now consider the mapping $\Psi_f : R/H \longrightarrow S$, defined by $\Psi_f(x+H) = f(x)$. By Theorem 3.3.8, Ψ_f is a bijection so the proof will be complete once we prove that Ψ_f is a homomorphism. We have

$$\begin{aligned}\Psi_f((x+H)+(y+H)) &= \Psi_f(x+y+H) = f(x+y) = f(x)+f(y)\\ &= \Psi_f(x+H)+\Psi_f(y+H),\end{aligned}$$

and

$$\begin{aligned}\Psi_f((x+H)(y+H)) &= \Psi_f(xy+H) = f(xy) = f(x)f(y)\\ &= \Psi_f(x+H)\Psi_f(y+H).\end{aligned}$$

The First Isomorphism Theorem now follows.

Theorem 4.4.6. *(The First Isomorphism Theorem, version 2). Let R, S be rings and let $f : R \longrightarrow S$ be a homomorphism. Then $R/\mathbf{Ker}f \cong \mathbf{Im}f$, a subring of S.*

Proof. The restriction of f to the mapping $R \longrightarrow \mathbf{Im}f$ is an epimorphism and hence, from Theorem 4.4.5, we see that $\mathbf{Im}f \cong R/\mathbf{Ker}f$. Finally, by Proposition 4.4.3, $\mathbf{Im}f$ is a subring of S.

We now consider some applications of these results.

The characteristic of a ring

Let R be a ring and let $f : \mathbb{Z} \longrightarrow R$ be the mapping defined by $f(n) = n1_R$, where $n \in \mathbb{Z}$. Clearly,

$$\begin{aligned}f(n+k) &= (n+k)1_R = n1_R + k1_R = f(n)+f(k) \text{ and}\\ f(nk) &= (nk)1_R = n(k1_R) = (n1_R)(k1_R) = f(n)f(k)\end{aligned}$$

for all $n,k \in \mathbb{Z}$. By Proposition 4.4.3,

$$\mathbf{Im}f = \{n1_R \mid n \in \mathbb{Z}\} = \mathbb{Z}1_R$$

is a subring of R. By Theorem 4.4.6, $\mathbb{Z}1_R = \mathbf{Im}f \cong \mathbb{Z}/\mathbf{Ker}f$. By Proposition 4.4.3, $\mathbf{Ker}f$ is an ideal of $\mathbb{Z}$. From the description of ideals in the ring $\mathbb{Z}$, obtained in Section 4.3, we have $\mathbf{Ker}f = n\mathbb{Z}$, for some fixed, but arbitrary, $n \geq 0$.

If $\mathbf{Ker}f = \{0\}$, then $\mathbb{Z}1_R \cong \mathbb{Z}$. In this case we say that the ring R has characteristic 0 and write $\mathbf{char}R = 0$. If $\mathbf{Ker}f = n\mathbb{Z}$, where $n > 0$, then $\mathbb{Z}1_R \cong \mathbb{Z}/n\mathbb{Z}$ and in this case we say that the ring R has characteristic $n > 0$ and write $\mathbf{char}R = n$. In this case if a is an arbitrary element of R, then

$$na = (n1_R)a = 0_R a = 0_R.$$

If n is not prime, then $n = kt$, where $1 < k,t < n$. Certainly $k,t \notin n\mathbb{Z}$. Therefore $k1_R \neq 0_R$ and $t1_R \neq 0_R$. However,

$$(k1_R)(t1_R) = (kt)1_R = n1_R = 0_R.$$

So if $\mathbf{char}R$ is not prime then the ring R has zero-divisors. Our argument gives us the proof of the following result.

Proposition 4.4.7. *If a ring R has no zero-divisors then either $\mathbf{char}R = 0$, or $\mathbf{char}R = p$, for some prime p. In particular, if R is a division ring or an integral domain, then either $\mathbf{char}R = 0$ or $\mathbf{char}R = p$, for some prime p.*

We now give some examples of homomorphisms. Let R be an integral domain and let K be a unitary subring of R. Let α be an element of R and consider the mapping $\vartheta : K[X] \longrightarrow R$, defined by $\vartheta(f(X)) = f(\alpha)$ for each polynomial $f(X) \in K[X]$. In Section 4.2 we noted that $(f+g)(\alpha) = f(\alpha) + g(\alpha)$ and $(fg)(\alpha) = f(\alpha)g(\alpha)$, which together imply

$$\vartheta(f+g) = \vartheta(f) + \vartheta(g) \text{ and } \vartheta(fg) = \vartheta(f)\vartheta(g).$$

Thus the mapping ϑ is a homomorphism. By Proposition 4.4.3 we deduce that

$$\mathbf{Im}(\vartheta) = \{f(\alpha) | f(X) \in K[X]\} = K[\alpha]$$

is a subring of R, and

$$\mathbf{Ker}(\vartheta) = \{f(X) | f(\alpha) = 0_R\} = \mathbf{Ann}_{K[X]}(\alpha)$$

is an ideal of $K[X]$. Furthermore, Theorem 4.4.6 shows that

$$K[\alpha] \cong K[X]/\mathbf{Ann}_{K[X]}(\alpha).$$

Definition 4.4.8. *Let R be an integral domain and let K be a unitary subring of R. An element $\alpha \in R$ is called algebraic over K, if $\mathbf{Ann}_{K[X]}(\alpha)$ is a nonzero ideal. Otherwise we will say that α is transcendental over K.*

Thus α is algebraic over K precisely when there exists a polynomial $f(X) \in K[X]$ which has α as a root, and α is transcendental over K when there is no such polynomial. A real number β is called an algebraic number, if β is algebraic over $\mathbb{Q}$. Otherwise we will say that β is a transcendental number. We note that if α is transcendental over K, then $K[\alpha] \cong K[X]$.

It is easy to find examples of algebraic numbers. For example, $2, \sqrt{3}$ are algebraic. However it is not easy to prove that a given number is transcendental. In the nineteenth century F. Lindemann proved that e is a transcendental number. Also in the nineteenth century, J. Liouville proved that π is transcendental. The Russian mathematician, A. O. Gelfond (1906–1968) developed an advanced theory allowing us to determine the transcendency of a wide class of numbers appearing in mathematical analysis. Many of the proofs of these facts use powerful analytic methods so we do not consider them here.

In Section 4.1 we proved that the subset

$$\mathbb{Z}[\sqrt{r}] = \{a + b\sqrt{r} \mid a, b \in \mathbb{Z}\}$$

is a subring of $\mathbb{C}$, where r is an integer with the property that $\sqrt{r} \notin \mathbb{Q}$. We construct next a matrix copy of this ring. Consider the mapping $f : \mathbb{Z}[\sqrt{r}] \longrightarrow M_2(\mathbb{Z})$, defined by

$$f(a + b\sqrt{r}) = \begin{pmatrix} a & b \\ rb & a \end{pmatrix}, \text{ whenever } a, b \in \mathbb{Z}.$$

If $\alpha = a + b\sqrt{r}, \beta = c + d\sqrt{r}$ are arbitrary elements of $\mathbb{Z}[\sqrt{r}]$, then

$$\alpha + \beta = (a + b\sqrt{r}) + (c + d\sqrt{r}) = (a + c) + \sqrt{r}(b + d),$$
$$\alpha\beta = (ac + bdr) + \sqrt{r}(ad + bc).$$

On the other hand,

$$\begin{pmatrix} a & b \\ rb & a \end{pmatrix} + \begin{pmatrix} c & d \\ rd & c \end{pmatrix} = \begin{pmatrix} a+c & b+d \\ r(b+d) & a+c \end{pmatrix}, \quad \text{and}$$

$$\begin{pmatrix} a & b \\ rb & a \end{pmatrix} \begin{pmatrix} c & d \\ rd & c \end{pmatrix} = \begin{pmatrix} ac+rbd & ad+bc \\ rbc+rad & ac+rbd \end{pmatrix},$$

so we see that $f(\alpha+\beta) = f(\alpha)+f(\beta)$ and $f(\alpha\beta) = f(\alpha)f(\beta)$. Hence f is a homomorphism. Suppose that $\alpha \in \mathbf{Ker}(f)$. By the definition of f, we see that $a = b = 0$, so that $\mathbf{Ker}(f) = \{0\}$. By Theorem 4.4.4, f is a monomorphism, so $\mathbb{Z}[\sqrt{r}] \cong \mathbf{Im}(f)$.

Finally we consider one more example from the ring of matrices. For the sake of simplicity, we will consider 3×3 matrices. In Section 4.3, we proved that $\mathbf{NT}_n(\mathbb{R})$ is an ideal of $\mathbf{T}_n(\mathbb{R})$, and $\mathbf{D}_n(\mathbb{R})$ is a subring of $\mathbf{T}_n(\mathbb{R})$. Consider the mapping $f : \mathbf{T}_3(\mathbb{R}) \longrightarrow \mathbf{D}_3(\mathbb{R})$ defined by

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix} \xrightarrow{f} \begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}.$$

Using properties of diagonal elements of triangular matrices, we can easily deduce that f is an epimorphism. Clearly, $\mathbf{Ker}(f) = \mathbf{NT}_3(\mathbb{R})$, and Theorem 4.4.5 shows that $\mathbf{T}_3(\mathbb{R})/\mathbf{NT}_3(\mathbb{R}) \cong \mathbf{D}_3(\mathbb{R})$.

We note, more generally that $\mathbf{T}_n(\mathbb{R})/\mathbf{NT}_n(\mathbb{R}) \cong \mathbf{D}_n(\mathbb{R})$.

Quite often in Algebra it is useful, when given an algebraic object A, to think of A as being contained in some larger algebraic object B of the same type, which we may have to construct. The concept of monomorphism plays a significant role in such constructions which are related to the idea of an extension of an algebraic object. In constructions of this kind, there is usually some subtlety that is not often emphasized in beginning textbooks. In reality the object B may not contain A precisely, but an isomorphic copy of A, and then it is said that the object A is identified with its isomorphic copy and that isomorphic objects in algebra do not differ. This point is often very difficult to understand. One example when this might occur is when A is a set of numbers, having an isomorphic image consisting of a set of matrices, and the extension takes place in the set of matrices. Therefore we present the following theorem which indicates how to build the extension of the strictly algebraic object A, taking advantage of the presence of a monomorphism of the object into some other object. This theorem is very technical, and its proof can be skipped.

Theorem 4.4.9. *(Theorem on ring extensions). Let R, K be rings and let $f : R \longrightarrow K$ be a monomorphism. Then there exists a ring S such that S is*

isomorphic to K and R is a subring of S. If $1_R, 1_K$ *are the identity elements of the rings R and K respectively and* $f(1_R) = 1_K$*, then* 1_R *is the identity element of S.*

Proof. We may assume that $K \cap R = \emptyset$. Indeed, if this is not true we can replace K by an isomorphic image having empty intersection with R. For example, we can put $K_1 = K \times \{0\}$ and define operations by the rules

$$(x,0) + (y,0) = (x+y,0) \text{ and } (x,0)(y,0) = (xy,0).$$

Let $U = \mathbf{Im}f$. By Proposition 4.4.3, U is a subring of K and $U \cong R$. Let $A = K \setminus U$ and $S = R \cup A$. Let $g : S \longrightarrow K$ be the mapping defined by

$$g(x) = \begin{cases} f(x), \text{ if } x \in R, \\ x, \text{ if } x \in A. \end{cases}$$

It is easy to check that g is a bijection from S onto K. We now define operations of addition, $\oplus$, and multiplication, $\otimes$, on S by

$$x \oplus y = g^{-1}(g(x) + g(y)) \text{ and } x \otimes y = g^{-1}(g(x)g(y))$$

for arbitrary $x, y \in S$. Then,

$$\begin{aligned} g(x \oplus y) &= g(g^{-1}(g(x) + g(y))) = g(x) + g(y) \text{ and} \\ g(x \otimes y) &= g(g^{-1}(g(x)g(y))) = g(x)g(y). \end{aligned}$$

We have to show that S is a ring and to this end we have

$$x \oplus y = g^{-1}(g(x) + g(y)) = g^{-1}(g(y) + g(x)) = y \oplus x,$$

so $\oplus$ is commutative. Also

$$\begin{aligned} (x \oplus y) \oplus z &= g^{-1}\big(g(x \oplus y) + g(z)\big) = g^{-1}\big((g(x) + g(y)) + g(z)\big) \\ &= g^{-1}\big(g(x) + (g(y) + g(z))\big) = g^{-1}\big(g(x) + g(y \oplus z)\big) \\ &= x \oplus (y \oplus z), \end{aligned}$$

so $\oplus$ is associative. Further, if $x \in R$ then,

$$x \oplus 0_R = g^{-1}\big(g(x) + g(0_R)\big) = g^{-1}(g(x) + 0_K) = g^{-1}(g(x)) = x.$$

Also, if $x \in R$ then

$$x \oplus (-x) = g^{-1}\big(g(x) + g(-x)\big) = g^{-1}\big(f(x + (-x))\big) = g^{-1}\big(f(0_R)\big) = 0_R.$$

If $x \in A$, then $-x \notin U$, because U is a subring. Thus

$$x \oplus (-x) = g^{-1}\big(g(x) + g(-x)\big) = g^{-1}\big(x + (-x)\big) = g^{-1}(0_K) = 0_R$$

and it follows that S is an abelian group under the operation $\oplus$.

Moreover,

$$\begin{aligned}(x \oplus y) \otimes z &= g^{-1}\big(g(x \oplus y)g(z)\big) = g^{-1}\big((g(x) + g(y))g(z)\big) \\ &= g^{-1}\big(g(x)g(z) + g(y)g(z)\big) = g^{-1}\big(g(x \otimes z) + g(y \otimes z)\big) \\ &= (x \otimes z) \oplus (y \otimes z)\end{aligned}$$

We can prove the equation

$$x \otimes (y \oplus z) = (x \otimes y) \oplus (x \otimes z)$$

in a similar fashion. Also, $\otimes$ is associative since

$$\begin{aligned}(x \otimes y) \otimes z &= g^{-1}\big(g(x \otimes y)g(z)\big) = g^{-1}\big((g(x)g(y))g(z)\big) \\ &= g^{-1}\big(g(x)(g(y)g(z))\big) = g^{-1}\big(g(x)g(y \otimes z)\big) = x \otimes (y \otimes z).\end{aligned}$$

Furthermore,

$$x \otimes g^{-1}(1_K) = g^{-1}\big(g(x)g(g^{-1}(1_K))\big) = g^{-1}(g(x)1_K) = g^{-1}\big(g(x)\big) = x$$

and

$$g^{-1}(1_K) \otimes x = g^{-1}\big(g(g^{-1}(1_K))g(x)\big) = g^{-1}\big(1_K g(x)\big) = g^{-1}\big(g(x)\big) = x,$$

so that $g^{-1}(1_K)$ is the identity element of S. In particular, if $f(1_R) = g(1_R) = 1_K$, then 1_R is the identity element of S. It follows that S is a ring. We have already proved that

$$g(x \oplus y) = g(x) + g(y) \text{ and } g(x \otimes y) = g(x)g(y),$$

which shows that g is an isomorphism.

If $x, y \in R$, then

$$\begin{aligned} x \oplus y &= g^{-1}\big(g(x)+g(y)\big) = g^{-1}\big(f(x)+f(y)\big) \\ &= g^{-1}\big(f(x+y)\big) = g^{-1}\big(g(x+y)\big) = x+y \text{ and} \\ x \otimes y &= g^{-1}\big(g(x)g(y)\big) = g^{-1}\big(f(x)f(y)\big) = g^{-1}\big(f(xy)\big) \\ &= g^{-1}\big(g(xy)\big) = xy \end{aligned}$$

Hence the restriction of these operations to R give rise to the original operations on R. Thus, by Theorem 4.1.4, R is a subring of S.

This theorem shows that to construct an extension E of the ring R it is sufficient to construct a monomorphism of R into some ring E isomorphic to the ring K, having the required set of properties.

Exercise Set 4.4

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

4.4.1. On the set $\mathbb{R} \times \mathbb{R}$ we define the operations of addition and multiplication by the following rules: $(a,b,)+(a_1,b_1,) = (a+a_1, b+b_1,)$, $(a,b)(a_1,b_1) = (aa_1 - bb_1, ab_1 + ba_1)$. Is $\mathbb{R} \times \mathbb{R}$ a ring? If the answer is yes, is there a monomorphism from the field of complex numbers into this ring?

4.4.2. Let $K = \mathbb{Z}[\sqrt{2}] = \{a+b\sqrt{2} \mid a,b \in \mathbb{Z}\}$. Let L be the subset of $\mathbf{M}_2(\mathbb{Z})$ consisting of all matrices of the form $\begin{pmatrix} a & b \\ 2b & a \end{pmatrix}$. Prove that K is a subring of $\mathbb{R}$, L is a subring of $\mathbf{M}_2(\mathbb{Z})$, and that K and L are isomorphic.

4.4.3. Let $K = \mathbb{Z}[\sqrt{3}] = \{a+b\sqrt{3} \mid a,b \in \mathbb{Z}\}$, and let L be the subset of $\mathbf{M}_2(\mathbb{Z})$ consisting of all matrices of the form $\begin{pmatrix} a & b \\ 3b & a \end{pmatrix}$. Prove that K is a subring of $\mathbb{R}$, L is a subring of $\mathbf{M}_2(\mathbb{Z})$, and that K and L are isomorphic.

4.4.4. Let $P_1 = \{x+y\sqrt{2} \mid x,y \in \mathbb{Q}\}, P_2 = \{x+y\sqrt{5} \mid x,y \in \mathbb{Q}\}$. Prove that P_1 and P_2 are subrings of $\mathbb{R}$. Is the map $f : P_1 \longrightarrow P_2$ defined by $f(x+y\sqrt{2}) = x+y\sqrt{5}$ an isomorphism from P_1 toP_2?

4.4.5. Let $P_1 = \{x+y\sqrt{3} \mid x,y \in \mathbb{Q}\}, P_2 = \{x+y\sqrt{7} \mid x,y \in \mathbb{Q}\}$. Prove that P_1 and P_2 are subrings of $\mathbb{R}$. Is the map $f : P_1 \longrightarrow P_2$ defined by the rule $f(x+y\sqrt{3}) = x+y\sqrt{7}$ an isomorphism from P_1 to P_2?

4.4.6. Let $K = \mathbb{Z}[i\sqrt{3}] = \{a+ib\sqrt{3} \mid a,b \in \mathbb{Z}\}$. Prove that K is a subring of $\mathbb{C}$. Define the mapping $f : K \longrightarrow \mathbf{M}_2(\mathbb{Z})$ by the rule $f(a+ib\sqrt{3}) = \begin{pmatrix} a & -3b \\ b & a \end{pmatrix}$. Is f a monomorphism?

4.4.7. Let $K = \mathbb{Z}[i\sqrt{2}] = \{a+ib\sqrt{2} \mid a,b \in \mathbb{Z}\}$. Prove that K is a subring of $\mathbb{C}$. Define the mapping $f : K \longrightarrow \mathbf{M}_2(\mathbb{Z})$ by the rule $f(a+ib\sqrt{2}) = \begin{pmatrix} a & -2b \\ b & a \end{pmatrix}$. Is f a monomorphism?

4.4.8. Let $\mathbb{Q}[\sqrt{p}] = \{x+y\sqrt{p} \mid x,y \in \mathbb{Q}\}$ where p is a prime. Is $\mathbb{Q}[\sqrt{p}]$ a subring of $\mathbb{R}$? Are $\mathbb{Q}[\sqrt{7}]$ and $\mathbb{Q}[\sqrt{5}]$ isomorphic?

4.4.9. Consider the quotient ring $\mathbb{Z}/3\mathbb{Z}$. What is the order of this quotient ring? What can we say regarding its additive group?

4.4.10. Consider the quotient ring $\mathbb{Z}/4\mathbb{Z}$. What is the order of this quotient ring? What can we say regarding its additive group?

4.4.11. Decide for which integers m,n we have $m\mathbb{Z} \cong n\mathbb{Z}$, as rings.

4.4.12. Let R be a commutative ring. Prove that R is a field if and only if every epimorphism $f : R \longrightarrow S$, where S is a nonzero ring, is an isomorphism.

4.4.13. Let R be a ring and let I be an ideal of R. Prove that there is a natural homomorphism of rings $f : \mathbf{M}_2(R) \longrightarrow \mathbf{M}_2(R/I)$, and find the kernel of this homomorphism. Deduce that $\mathbf{M}_2(R)/\mathbf{M}_2(I) \cong \mathbf{M}_2(R/I)$.

4.4.14. Prove the Second Isomorphism Theorem for Rings, which states that if R is a ring (not necessarily with identity), I is an ideal of R and S is a subring of R then $S+I/I \cong S/(S \cap I)$.

4.4.15. Prove the Third Isomorphism Theorem for Rings, which states that if R is a ring (not necessarily with identity) and I,J are ideals of R such that $I \leq J$ then $(R/I)/(J/I) \cong R/J$.

4.4.16. Let R be a ring not necessarily with an identity element. Let $K = R \times \mathbb{Z}$. Define operations on K by

$$(x,n)+(y,k) = (x+y,n+k) \text{ and } (x,n)(y,k) = (xy+kx+ny,nk)$$

for all $x,y \in R$ and $n,k \in \mathbb{Z}$. Prove that K is a ring with identity element.

4.4.17. Let R be a ring, not necessarily with multiplicative identity. Prove that there exists a ring K with multiplicative identity such that R is isomorphic to a subring of K.

4.4.18. Let m,n be relatively prime integers. Prove that $\mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/mn\mathbb{Z}$ by first showing that the function f defined by $f(a) = (a + m\mathbb{Z}, a + n\mathbb{Z})$, for all $a \in \mathbb{Z}$, is a ring homomorphism. Then use the First Isomorphism Theorem.

4.4.19. Let R,S be rings, not necessarily with identity, suppose that S has at least two elements and let $f : R \longrightarrow S$ be a surjective ring homomorphism. If R does have an identity element then prove that S also has an identity element.

4.4.20. Let $R = \{a + bi | a, b \in \mathbb{Z}\}$ and let $I = 3R$. Find the order of the quotient ring R/I and prove that R/I is a field. Is $R/5R$ also a field?

5

FIELDS

5.1 FIELDS: BASIC PROPERTIES AND EXAMPLES

In this chapter we consider one of the main concepts in algebra, namely the concept of a field. As we saw in Section 4.1 fields are rings in which the additive and multiplicative structures have similar properties. To be more specific, in a field F, not only is the set F an abelian group under addition but also the set of nonzero elements of F is an abelian group with respect to multiplication. Thus in a field not only are addition, subtraction, and multiplication possible, but also division can be defined, since we can find multiplicative inverses of nonzero elements. The presence of such strong conditions has made it possible to develop a very deep theory of fields. Currently, field theory is one of the most advanced algebraic theories, having a variety of connections with other areas of mathematics. To highlight the importance of fields, we devote this final chapter to the study of some of their elementary properties.

In Section 4.1, we have already given some examples of fields. But in order to expand the array of examples, we first introduce the concept of a subfield, and then look at examples of fields and subfields.

Definition 5.1.1. *Let F be a field. A subset P of F is called a subfield if P is a unitary subring of F and $P \setminus \{0_F\}$ is a subgroup of* $\mathbf{U}(F)$.

An Introduction to Essential Algebraic Structures, First Edition. Martyn R. Dixon, Leonid A. Kurdachenko and Igor Ya. Subbotin.

Thus a subfield P of F is a subset, in which the field operations of addition and multiplication are closed; furthermore P must contain the identity element of F and P must also be a field under the restriction of these operations to P. If P is a subfield of a field F, then we say also that F is a field extension of P, or simply that F is an extension of P.

Theorem 5.1.2. *Let F be a field. The subset P of F is a subfield if and only if it satisfies the following conditions:*

(**SF 1**) *if $x, y \in P$, then $x - y \in P$;*
(**SF 2**) *if $x, y \in P$, then $xy \in P$;*
(**SF 3**) *the identity element 1_F of F belongs to P;*
(**SF 4**) *if x is a nonzero element of P, then $x^{-1} \in P$.*

Proof. Let P be a subfield of F. Since P is a unitary subring, it satisfies conditions (**SF 1**) – (**SF 3**) by Theorem 4.1.4. Clearly, 1_F is the identity element of P. Since $P \setminus \{0_F\}$ is a subgroup of $\mathbf{U}(F)$, Theorem 3.1.4 shows that P satisfies condition (**SF 4**).

Conversely, suppose that P satisfies conditions (**SF 1**) – (**SF 4**). Condition (**SF 3**) implies that P is nonempty. Then Theorem 4.1.4 shows that P is a subring of F. Moreover, by (**SF 3**), P is a unitary subring. Conditions (**SF 2**) and (**SF 4**) together with Theorem 3.1.4 shows that $P \setminus \{0_F\}$ is a subgroup of $\mathbf{U}(F)$. Hence P is a subfield of F.

Corollary 5.1.3. *Let F be a field and let $\mathfrak{S}$ be a family of subfields of F. The intersection $\cap\mathfrak{S}$ of all subfields from this family is also a subfield of F.*

Proof. Let $V = \cap\mathfrak{S}$. Since every subfield contains 0_F and 1_F, we have 0_F, $1_F \in V$. Thus V satisfies (**SF 3**). Let $x, y \in V$. If U is an arbitrary element of $\mathfrak{S}$, then $x - y, xy \in U$, by **(SF 1)** and **(SF 2)**. Therefore $x - y, xy$ lie in the intersection of all elements of $\mathfrak{S}$. However this intersection is V, so we have $x - y, xy \in V$, which shows that V satisfies (**SF 1**) and (**SF 2**). In the same way we can prove that V satisfies (**SF 4**) and Theorem 5.1.2 implies that V is a subfield of F.

Definition 5.1.4. *Let F be a field. Then the intersection F_0 of all subfields of F is called the prime subfield of F. A field F is called prime if F coincides with its prime subfield.*

It is clear from this definition that if F is a prime field then F has no proper subfields. We now give some examples of fields.

1. The sets $\mathbb{C}$ of all complex numbers, the set $\mathbb{R}$ of all real numbers, and the set $\mathbb{Q}$ of all rational numbers are fields. Furthermore, $\mathbb{R}$ is a subfield of $\mathbb{C}$ and $\mathbb{Q}$ is a subfield of $\mathbb{R}$.
2. Also, $\mathbb{Q}$ is a prime field. To see this, let P be some subfield of $\mathbb{Q}$. From Theorem 5.1.2 it follows that $1 \in P$. By (**SF 1**) we have $2 = 1+1 \in P$, $3 = 2+1 \in P$, and similarly, for each $n \in \mathbb{N}$ we have $n = n \cdot 1 \in P$. Again by (**SF 1**) we see than $-n = 0 - n \in P$ for each $n \in \mathbb{N}$, so that $n \in P$ for each $n \in \mathbb{Z}$. If $0 \neq k \in \mathbb{Z}$, then by (**SF 4**) $\frac{1}{k} \in P$. Now, for all $r, k \in \mathbb{Z}$, where $k \neq 0$, we have

$$\frac{r}{k} = r\left(\frac{1}{k}\right) \in P.$$

Thus $\mathbb{Q} \subseteq P$ so this shows that $P = \mathbb{Q}$.

The field $\mathbb{R}$ is not prime since it strictly contains the field $\mathbb{Q}$. Likewise, $\mathbb{C}$ is not prime, since it strictly contains $\mathbb{R}$. Indeed, between $\mathbb{Q}$ and $\mathbb{R}$ there are many subfields and we now consider some of these.

3. Let r be an integer and suppose that $\sqrt{r} \notin \mathbb{Q}$. Put

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

Let α, β be arbitrary elements of $\mathbb{Q}(\sqrt{r})$, say $\alpha = a + b\sqrt{r}$ and $\beta = a_1 + b_1\sqrt{r}$. Easy computations show that

$$\alpha - \beta = (a - a_1) + (b - b_1)\sqrt{r} \text{ and } \alpha\beta = (aa_1 + bb_1 r) + (ab_1 + ba_1)\sqrt{r}.$$

It follows that $\alpha - \beta, \alpha\beta \in \mathbb{Q}(\sqrt{r})$. Clearly, $1 \in \mathbb{Q}(\sqrt{r})$. Also if $\alpha \neq 0$, then $a^2 - rb^2 \neq 0$ since $\sqrt{r} \notin \mathbb{Q}$ so

$$\gamma = \frac{a}{a^2 - rb^2} + \left(\frac{-b}{a^2 - rb^2}\right)\sqrt{r} \in \mathbb{Q}(\sqrt{r}).$$

By a direct computation it is easy to verify that $\alpha\gamma = \gamma\alpha = 1$ so $\gamma = \alpha^{-1}$. Then Theorem 5.1.2 shows that $\mathbb{Q}(\sqrt{r})$ is a subfield of $\mathbb{C}$ called a *quadratic extension of* $\mathbb{Q}$. Of course, when $r > 0$ then $\mathbb{Q}(\sqrt{r})$ is a subfield of $\mathbb{R}$.

Proposition 4.3.14, serves as another source for the construction of examples.

4. Let p be a prime, then the quotient ring $\mathbb{Z}/p\mathbb{Z}$ is a field by Proposition 4.3.11. Let $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. This is a finite field with p elements. Furthermore,

$$\mathbb{F}_p = \{p\mathbb{Z} = 0 + p\mathbb{Z}, 1 + p\mathbb{Z}, 2 + p\mathbb{Z}, \ldots, p - 1 + p\mathbb{Z}\}.$$

Let $\mathbf{j} = j + p\mathbb{Z}$ and $\mathbf{k} = k + p\mathbb{Z}$, for $0 \le j, k \le p-1$. (We shall drop this special notation for the elements of $\mathbb{F}_p$ rather quickly, so that $\mathbf{j}$ will soon be written as just j, a symbol which may then mean either an integer or an element of $\mathbb{F}_p$.) We have

$$\mathbf{j} + \mathbf{k} = (j + p\mathbb{Z}) + (k + p\mathbb{Z}) = j + k + p\mathbb{Z}.$$

Suppose that $j + k \ge p$. By Theorem 2.2.1, there exist positive integers b, r such that $j + k = bp + r$, where $0 < r < p$. In this case, $bp + r + p\mathbb{Z} = r + p\mathbb{Z}$. Similarly, if $jk \ge p$, then again, using Theorem 2.2.1, we obtain the decomposition $jk = cp + u$, where $0 < u < p$, and $cp + u + p\mathbb{Z} = u + p\mathbb{Z}$. Thus we obtain the following simple rules for addition and multiplication of elements in $\mathbb{F}_p$. The sum (respectively product) $\mathbf{j} + \mathbf{k}$ (respectively $\mathbf{jk}$) of elements $\mathbf{j}, \mathbf{k} \in \mathbb{F}_p$ is the remainder upon division of $j + k$ (respectively jk) by the prime p. We next give sample addition and multiplication tables for the elements of $\mathbb{F}_p$ for the primes $p = 2, 3, 5$.

$\mathbb{F}_2$

+	0	1
0	0	1
1	1	0

,

·	0	1
0	0	0
1	0	1

.

$\mathbb{F}_3$

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

,

·	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

.

$\mathbb{F}_5$

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

,

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

.

We note that for each prime p, $\mathbb{F}_p$ is a prime field. In fact, if P is a subfield of $\mathbb{F}_p$, then P is a subgroup, under addition, of the additive group of $\mathbb{F}_p$. Theorem 3.2.8 shows that the order, $|P|$, of P is a divisor of $\left|\mathbb{F}_p\right| = p$. Since P is a subfield, P has at least two elements and it follows that $|P| = p$. Consequently $P = \mathbb{F}_p$, so $\mathbb{F}_p$ has no proper subfields.

5. Next we complicate the above construction by a small amount. First we consider the polynomial ring $\mathbb{F}_2[X]$. The polynomial $d(X) = X^2 + X + 1$ has

no roots in $\mathbb{F}_2$, since $d(0) = d(1) = 1$, and hence Corollary 4.2.6 shows that $d(X)$ is irreducible over $\mathbb{F}_2$. Proposition 4.3.17 implies that the ideal $M = d(X)\mathbb{F}_2[X]$ is maximal, and Proposition 4.3.16 shows that the quotient ring $F = \mathbb{F}_2[X]/M$ is a field. We discuss this field in some detail. Let $f(X)$ be an arbitrary polynomial with coefficients in $\mathbb{F}_2$. By Theorem 4.2.3 $f(X) = q(X)d(X) + r(X)$ where $q(X), r(X) \in \mathbb{F}_2[X]$ and either $r(X) = 0$ or $\mathbf{deg}\, r(X) < 2$. We have $f(X) + M = r(X) + M$, which shows that the field F actually consists of the following elements:

$$\theta = 0 + M, \varepsilon = 1 + M, \kappa = X + M, \lambda = X + 1 + M.$$

We next construct the addition and multiplication tables for these cosets. For addition this is an easy task, while for multiplication we need to remember that $d(X) = X^2 + X + 1$ which can be rewritten as $X^2 = d(X) + X + 1$ since, in $\mathbb{F}_2$, we have $-1 = 1$. Hence

$$\kappa^2 = X^2 + M = d(X) + X + 1 + M = (X + 1) + M = \lambda.$$

Since

$$(X + 1)^2 = X^2 + 2X + 1 = X^2 + 1 = d(X) + X,$$

we have

$$\lambda^2 = (X + 1)^2 + M = X + M = \kappa;$$

$$\text{Also } (X + 1)X = X^2 + X = d(X) + 1,$$

so that

$$\lambda\kappa = (X + 1)X + M = 1 + M = \varepsilon.$$

The addition and multiplication table for $F = \mathbb{F}_4$ can now be constructed as follows.

$+$	θ	ε	κ	λ
θ	θ	ε	κ	λ
ε	ε	θ	λ	κ
κ	κ	λ	θ	ε
λ	λ	κ	ε	θ

,

$\cdot$	θ	ε	κ	λ
θ	θ	θ	θ	θ
ε	θ	ε	κ	λ
κ	θ	κ	θ	ε
λ	θ	λ	ε	κ

.

In Section 4.4 we defined the characteristic of a ring. Since a field has no zero divisors, Proposition 4.4.7 shows that either $\mathbf{char}(F) = 0$ or $\mathbf{char}(F) = p$ for some prime p. For the examples discussed here, it is easy to see that

the fields $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are fields of characteristic 0. On the other hand, for the prime p, $p \cdot \mathbf{1} = \mathbf{0}$, so $\mathbf{char}(\mathbb{F}_p) = p$ and similarly $\mathbf{char}(\mathbb{F}_4) = 2$.

Our next goal is to characterize the prime subfield for every field. Before doing so, we consider homomorphisms of fields.

Let F, P be fields. Then the mapping $\theta : F \longrightarrow P$, defined by $\theta(x) = 0_P$, for each $x \in F$ is clearly a ring homomorphism. This homomorphism is called the zero homomorphism.

Theorem 5.1.5. *(Theorem on homomorphism of fields) Let F, P be fields and let $f : F \longrightarrow P$ be a nonzero (ring) homomorphism.*

(i) *If 1_F and 1_P are the identity elements of F and P respectively, then $f(1_F) = 1_P$;*
(ii) *The function f is a monomorphism and hence the field F is isomorphic to some subfield of P.*

Proof. Let $f(1_F) = u$. If $u = 0_P$, then

$$f(a) = f(a1_F) = f(a)f(1_F) = f(a)0_P = 0_P$$

for every element a of P. Thus f is the zero homomorphism and we obtain a contradiction, which shows that $u \neq 0_P$. However we have

$$u^2 = uu = f(1_F)f(1_F) = f(1_F^2) = f(1_F) = u.$$

Since u is nonzero, it has a multiplicative inverse, so that

$$1_P = uu^{-1} = u^2u^{-1} = u = f(1_F).$$

By Proposition 4.4.3, $\mathbf{Ker}(f)$ is an ideal of P. Since f is nonzero, Corollary 4.3.6 implies that $\mathbf{Ker}(f)$ is zero. Then Theorem 4.4.4 implies that f is a monomorphism, so that F is isomorphic to $\mathbf{Im}(f)$. Finally, Proposition 4.4.3 shows that $\mathbf{Im}(f)$ is a subring of P, so it satisfies (**SF 1**) and (**SF 2**). We have proved already that $\mathbf{Im}(f)$ satisfies (**SF 3**). If v is a nonzero element of $\mathbf{Im}(f)$, then $v = f(b)$ for some element $b \in F$. Clearly b is nonzero, so b has a multiplicative inverse. We have

$$1_P = f(1_F) = f(bb^{-1}) = f(b)f(b^{-1}) = vf(b^{-1}).$$

It follows that $v^{-1} = f(b^{-1}) \in \mathbf{Im}(f)$, so that $\mathbf{Im}(f)$ satisfies (**SF 4**). Theorem 5.1.2 proves that $\mathbf{Im}(f)$ is a subfield of P. This completes the proof.

We will use this theorem to describe all prime subfields and hence all prime fields. So far we have two types of example of prime fields, namely $\mathbb{Q}$ and $\mathbb{F}_p$, for every prime p. Our next theorem shows that there are essentially no other prime fields.

Theorem 5.1.6. *Let F be a field and let F_0 be the prime subfield of F.*

(i) *If* $\mathbf{char}F = p$ *is a prime then* $F_0 \cong \mathbb{F}_p$;
(ii) *If* $\mathbf{char}F = 0$, *then* $F_0 \cong \mathbb{Q}$.

Proof. Consider the mapping $f : \mathbb{Z} \longrightarrow F$ defined by $f(n) = n1_F$, where $n \in \mathbb{Z}$. As we proved above, this mapping is a homomorphism. Moreover, f is a nonzero homomorphism, because $0_F \neq 1_F = f(1)$. If $\mathbf{char}F = p$ is a prime, then $p1_F = 0_f$ so $p \in \mathbf{Ker}f$. Since $\mathbf{Ker}f$ is an ideal of $\mathbb{Z}$ it follows that $p\mathbb{Z} \leq \mathbf{Ker}(f)$. In Section 4.3 we noted that $p\mathbb{Z}$ is a maximal ideal of $\mathbb{Z}$. Hence either $\mathbf{Ker}(f) = p\mathbb{Z}$ or $\mathbf{Ker}(f) = \mathbb{Z}$. In the latter case, f must be the zero homomorphism, contrary to the choice of f. Therefore $\mathbf{Ker}(f) = p\mathbb{Z}$. Let $\psi : \mathbb{Z}/p\mathbb{Z} \longrightarrow F$ be defined by $\psi(n + p\mathbb{Z}) = f(n)$, where $n \in \mathbb{Z}$. This is also a homomorphism and this homomorphism is nonzero, because $\psi(1 + p\mathbb{Z}) = f(1) = 1_F \neq 0_F$. Clearly, $\mathbf{Im}\,\psi = \mathbf{Im}f = \mathbb{Z}1_F$ and we deduce the result in this case by applying Theorem 5.1.5.

Suppose now that $\mathbf{char}F = 0$. In this case, $\mathbf{Ker}f = \{0_F\}$ and Theorem 4.4.4 implies that f is a monomorphism. Since $1_F \in F_0$ it follows that $\mathbb{Z}1_F$ is a subring of F_0 . We extend the mapping f to a mapping $f_1 : \mathbb{Q} \longrightarrow F$ in the following way. Let $\frac{m}{n}$ be an arbitrary element of $\mathbb{Q}$. Then $n \neq 0$ and hence $f(n) = n1_F \neq 0_F$. Since F is a field, its nonzero element $n1_F$ is invertible in F. We now set

$$f_1\left(\frac{m}{n}\right) = (m1_F)(n1_F)^{-1}.$$

The function f_1 is well-defined. For, if $\frac{k}{t} = \frac{m}{n}$ then we have

$$kn = tm \text{ so } (k1_F)(n1_F) = (t1_F)(m1_F)$$

and hence

$$(m1_F)(n1_F)^{-1} = (k1_F)(t1_F)^{-1} = (t1_F)^{-1}(k1_F).$$
$$\text{Thus } f_1(m/n) = (m1_F)(n1_F)^{-1} = (k1_F)(t1_F)^{-1} = f_1(k/t),$$

and that f_1 is well-defined now follows. The mapping f_1 is an extension of f since if $n \in \mathbb{Z}$, then $n = \frac{n}{1}$ so that

$$f_1(n) = f_1\left(\frac{n}{1}\right) = (n1_F)(1_F)^{-1} = (n1_F)(1_F)(1_F)^{-1} = n1_F = f(n).$$

Hence the mapping f_1 is nonzero. The mapping f_1 is a homomorphism since

$$\begin{aligned} f_1\left(\frac{m}{n} + \frac{k}{t}\right) &= f_1\left(\frac{mt+kn}{nt}\right) = \big((mt+kn)1_F\big)\big((nt)1_F\big)^{-1} \\ &= \big((mt)1_F + (kn)1_F\big)\big((n1_F)(t1_F)\big)^{-1} \\ &= \big((m1_F)(t1_F) + (k1_F)(n1_F)\big)(n1_F)^{-1}(t1_F)^{-1} \\ &= (m1_F)(t1_F)(n1_F)^{-1}(t1_F)^{-1} + (k1_F)(n1_F)(n1_F)^{-1}(t1_F)^{-1} \\ &= (m1_F)(n1_F)^{-1} + (k1_F)(t1_F)^{-1} = f_1\left(\frac{m}{n}\right) + f_1\left(\frac{k}{t}\right). \end{aligned}$$

Also

$$\begin{aligned} f_1\left(\left(\frac{m}{n}\right)\left(\frac{k}{t}\right)\right) &= f_1\left(\frac{mk}{nt}\right) = \big((mk)1_F\big)\big((nt)1_F\big)^{-1} \\ &= (m1_F)(k1_F)\big((n1_F)(t1_F)\big)^{-1} \\ &= (m1_F)(k1_F)(n1_F)^{-1}(t1_F)^{-1} \\ &= (m1_F)(n1_F)^{-1}(k1_F)(t1_F)^{-1} \\ &= f_1\left(\frac{m}{n}\right) f_1\left(\frac{k}{t}\right). \end{aligned}$$

By Theorem 5.1.5, $\mathbb{Q} \cong \mathbf{Im}f_1$. Since $\mathbf{Im}f_1$ is a subfield of F, it follows that $\mathbf{Im}f_1 \geq F_0$. On the other hand, $n1_F \in F_0$ for every $n \in \mathbb{Z}$ and if $n \neq 0$, then $(n1_F)^{-1} \in F_0$. Therefore $(m1_F)(n1_F)^{-1} \in F_0$ for every pair $m, n \in \mathbb{Z}$, where $n \neq 0$. Thus, $\mathbf{Im}f_1 \leq F_0$, and then $\mathbf{Im}f_1 = F_0$.

Corollary 5.1.7. *Let F be a prime field.*

(i) *If* $\mathbf{char}F = p$ *is a prime, then* $F \cong \mathbb{F}_p$;
(ii) *If* $\mathbf{char}F = 0$, *then* $F \cong \mathbb{Q}$.

The structure of the prime subfield of a field exerts a strong influence on the structure of the entire field.

Proposition 5.1.8. *Let F be a field.*

(i) *Suppose that* $\mathbf{char}(F) = 0$ *and that a is a nonzero element of F. If $na = ka$, for some $n, k \in \mathbb{Z}$, then $n = k$. In particular, if $na = 0_F$ then $n = 0$;*

(ii) *If* $\mathbf{char}(F) = p$ *is a prime, then $pa = 0_F$ for each element $a \in F$;*

(iii) *Suppose that* $\mathbf{char}(F) = p$ *is a prime and that a is a nonzero element of F. If $na = ka$, for some $n, k \in \mathbb{Z}$, then $n \equiv k \pmod{p}$. In particular, if $na = 0_F$, then p divides n;*

(iv) *If* $\mathbf{char}(F) = p$ *is prime, then $(a+b)^p = a^p + b^p$ for all elements $a, b \in F$;*

(v) *If* $\mathbf{char}(F) = p$ *is prime, then $(a-b)^p = a^p - b^p$ for all elements $a, b \in F$.*

Proof.

(i) If $\mathbf{char}(F) = 0$, then the mapping $f : \mathbb{Z} \longrightarrow F$ defined by $f(n) = n1_F$, is a monomorphism. It follows that $\mathbf{Ker}(f) = \{0\}$. Let $0_F \neq a \in F$ and suppose that $na = 0_F$. Clearly $na = (n1_F)a$. Since F has no zero-divisors, then $(n1_F)a = 0_F$ implies that $n1_F = 0_F$, so that $n \in \mathbf{Ker}(f) = \{0\}$. Suppose now that $na = ka$. Then $0_F = na - ka = (n-k)a$. By our work here we deduce that $n - k = 0$ or $n = k$.

(ii) We have $pa = (p1_F)a$. Since $p1_F = 0_F$ we deduce that $pa = \mathbf{0}_F$.

(iii) Let n be an integer such that $n1_F = 0_F$ and suppose that p does not divide n. Since p is a prime, $\mathbf{GCD}(n, p) = 1$ and Corollary 2.2.7 implies that there exist integers u, v such that $1 = un + vp$. Then

$$1_F = 1 \cdot 1_F = (un + vp)1_F = (un)1_F + (vp)1_F = u(n1_F) + v(p1_F) = 0_F,$$

and we obtain a contradiction. This contradiction shows that p is a divisor of n.

Let $0_F \neq a \in F$ and suppose that $na = 0_F$. Clearly $na = (n1_F)a$. Since F has no zero-divisors, then $(n1_F)a = 0_F$ implies that $n1_F = 0_F$. By our work here we deduce that p divides n in this case also. Now suppose that $na = ka$. Then $0_F = na - ka = (n-k)a$. We have already proved that in this case, p divides $n - k$, so that $n \equiv k \pmod{p}$.

(iv) In Section 2.1 we proved an equation of the form

$$(a+b)^p = a^p + \mathbf{C}_1^p a^{p-1}b + \mathbf{C}_2^p a^{p-2}b^2 + \cdots + \mathbf{C}_k^p a^{p-k}b^k + \ldots + \mathbf{C}_{p-1}^p ab^{p-1} + b^p,$$

where $\mathbf{C}_k^p = \frac{(p-k+1)\ldots(p-1)p}{k!}$, for $1 \leq k \leq p-1$. This formula holds in any commutative ring, so it holds in F. When $k < p$ and p is a prime,

$\mathbf{GCD}(\frac{(p-k+1)\ldots(p-1)}{k!}, p) = 1$, so that p divides $\mathbf{C}_k^p$ for $1 \leq k \leq p-1$. It follows that $\mathbf{C}_k^p a^{p-k} b^k = 0_F$ for $1 \leq k \leq p-1$. Hence $(a+b)^p = a^p + b^p$.

(v) If $p = 2$, then $2x = 0_F$, which shows that $-x = x$ for each $x \in F$. Then $a - b = a + b$ and $(a-b)^2 = (a+b)^2 = a^2 + b^2 = a^2 - b^2$.

Suppose now that p is an odd prime. Let $c = -b$, so that $a - b = a + c$. By (iv) $(a+c)^p = a^p + c^p$ and we have

$$c^p = (-b)^p = ((-1_F)b)^p = (-1_F)^p b^p = (-1_F)b^p = -b^p.$$

It follows that

$$(a-b)^p = (a+c)^p = a^p + c^p = a^p - b^p.$$

This completes the proof.

To conclude this section, we discuss a standard construction, which shows how fields arise from certain rings. This construction is based on the procedure for constructing rational numbers, well-known even from high school, and which we discussed in Section 2.4. If F is a field and R is a subring of F, then R has no zero-divisors. This condition is sufficient for us to be able to embed a commutative ring in field.

To this end, let R be a commutative ring (we often use the commutativity with little fanfare in what follows) with no zero-divisors, and let

$$V = \{(x, y) \mid x, y \in R, y \neq 0_R\}.$$

We define a partition $\mathfrak{D}$ on the set V as follows, and note that its definition is entirely analogous with the rule of equality of rational numbers. We will treat a fraction $\frac{x}{y}$ as some subset of V. Two pairs $(x, y), (u, v) \in V$ define one and the same fraction (i.e., they belong to the same subset from the family $\mathfrak{D}$) if and only if $xv = uy$. This defines the way we construct the family $\mathfrak{D}$. For every pair $(x, y) \in V$ we define the subset $\frac{x}{y}$ to be the subset consisting of all the pairs (u, v) such that $xv = uy$. Thus

$$\frac{x}{y} = \{(u, v) \in V | xv = uy\}.$$

We define the family $\mathfrak{D}$ to be the set whose elements are precisely the subsets $\frac{x}{y}$, for all pairs $(x, y) \in V$. We now show that every pair $(u, v) \in \frac{x}{y}$ satisfies $\frac{u}{v} = \frac{x}{y}$. To this end, let $(z, w) \in \frac{u}{v}$. Then

$$zv = wu \text{ and } xv = uy.$$

We multiply both sides of the first equality by y, and both sides of the second equality by w to obtain $zvy = wuy$ and $xvw = uyw$. Since R is commutative, $wuy = uyw$ so we have $zvy = xvw$ and hence, by commutativity again, $zyv = wxv$. Since R has no zero-divisors and $v \neq 0$ Proposition 4.1.9 implies that $zy = xw = wx$, so that $(z,w) \in \frac{x}{y}$. This proves that $\frac{u}{v} \subseteq \frac{x}{y}$, and a similar argument establishes $\frac{x}{y} \subseteq \frac{u}{v}$. Hence $\frac{u}{v} = \frac{x}{y}$. The subset $\frac{x}{y}$ is called a fraction with numerator x and denominator y.

The family $\mathfrak{D}$ of all fractions gives us a partition of V. Indeed, $\cup\mathfrak{D} = V$, because $(x,y) \in \frac{x}{y}$. Also, suppose that $(x_2,y_2) \in \frac{x}{y} \cap \frac{x_1}{y_1} \neq \emptyset$. As we proved above this implies that $\frac{x}{y} = \frac{x_2}{y_2}$ and $\frac{x_1}{y_1} = \frac{x_2}{y_2}$, so $\frac{x}{y} = \frac{x_1}{y_1}$. Thus fractions are either equal or disjoint. The relationship defining the elements of a fraction is an equivalence relation and the fractions themselves are the corresponding equivalence classes.

On $\mathfrak{D}$, we introduce operations of addition and multiplication in the same way as is done for rational numbers. Thus we let $\frac{x}{y} + \frac{u}{v} = \frac{xv+uy}{yv}$ and $(\frac{x}{y})(\frac{u}{v}) = \frac{xu}{yv}$, whenever $\frac{x}{y}, \frac{u}{v} \in \mathfrak{D}$.

Next we show that these definitions of addition and multiplication are well-defined, in that they do not depend upon the choice of elements used to define the fractions. To see this let $(x_1,y_1) \in \frac{x}{y}$ and $(u_1,v_1) \in \frac{u}{v}$. Then $xy_1 = yx_1$ and $uv_1 = vu_1$, so we obtain

$$\begin{aligned}(xv+uy)y_1v_1 &= xvy_1v_1 + uyy_1v_1 = xy_1vv_1 + uv_1yy_1\\ &= yx_1vv_1 + vu_1yy_1 = (x_1v_1 + u_1y_1)yv.\end{aligned}$$
$$\text{Also } (xu)(y_1v_1) = (xy_1)(uv_1) = (yx_1)(vu_1) = (x_1u_1)(yv).$$

It follows that

$$(x_1v_1 + u_1y_1, y_1v_1) \in \frac{x}{y} + \frac{u}{v} \text{ and } (x_1u_1, y_1v_1) \in \frac{x}{y}\frac{u}{v}$$

so the previous equations imply that

$$\frac{x_1}{y_1} + \frac{u_1}{v_1} = \frac{x}{y} + \frac{u}{v}, \text{ and } \frac{x_1}{y_1}\frac{u_1}{v_1} = \frac{x}{y}\frac{u}{v}.$$

Using the same method as used for rational numbers we can verify that the operations of addition and multiplication of fractions introduced above are commutative, associative, and connected by the distributive property. Addition has a zero element, which is the fraction $\frac{0_R}{y}$ (clearly $\frac{0_R}{y} = \frac{0_R}{v}$ for all nonzero elements $y, v \in R$). Multiplication has an identity element, which is the fraction $\frac{y}{y}$ (clearly $\frac{y}{y} = \frac{x}{x}$ for all nonzero elements $x, y \in R$). Finally, an additive inverse to the fraction $\frac{x}{y}$ is $\frac{-x}{y}$; a multiplicative inverse to the nonzero fraction $\frac{x}{y}$ is $\frac{y}{x}$ (since $\frac{x}{y} \neq \frac{0_R}{v}$, it follows that $x \neq 0_R$, and hence the fraction $\frac{y}{x}$ is defined).

This shows that the set $\mathfrak{D}$ is a field, called *the field of fractions* for the integral domain R.

We next define a mapping $f : R \longrightarrow \mathfrak{D}$. We note first that if $x \in R$ and y, v are arbitrary nonzero elements of R, then $\frac{xy}{y} = \frac{xv}{v}$. Accordingly, let $f(x) = \frac{xy}{y}$, where y here can be chosen arbitrarily in $R \setminus \{0\}$. We have $f(x+u) = \frac{(x+u)w}{w}$ and

$$f(x)+f(u) = \frac{xy}{y} + \frac{uv}{v} = \frac{xyv+uvy}{yv} = \frac{(x+u)(yv)}{yv},$$

so that $f(x+y) = f(x)+f(y)$. Similarly,

$$f(x)f(u) = \frac{xy}{y}\frac{uv}{v} = \frac{(xu)(yv)}{yv} = \frac{(xu)w}{w} = f(xu).$$

Hence f is a homomorphism. If $x \in \mathbf{Ker}(f)$, then $\frac{xy}{y} = \frac{0_R}{v}$, which implies that $xyv = y0_R = 0_R$. Since $yv \neq 0_R$, it follows that $x = 0_R$, so $\mathbf{Ker}(f) = \{0_R\}$. By Theorem 4.4.4, f is a monomorphism. We also remark that $f(1_R) = \frac{1_R y}{y} = \frac{y}{y}$ is the identity element of $\mathfrak{D}$. Theorem 4.4.9 shows that there exists a commutative ring $K \cong D$ such that R is a unitary subring of K.

Exercise Set 5.1

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

5.1.1. Let $P = \{x+y\sqrt{2} \mid x, y \in \mathbb{Q}\}$. Prove that P is a subfield of $\mathbb{R}$.

5.1.2. Let $P = \{x+y\sqrt{5} \mid x, y \in \mathbb{Q}\}$. Prove that P is a subfield of $\mathbb{R}$.

5.1.3. Let $P = \{x+y\sqrt{2} \mid x, y \in \mathbb{Q}\}$. In P solve the equation $x^2 - x - 3 = 0$.

5.1.4. Let $P = \{x+y\sqrt{5} \mid x, y \in \mathbb{Q}\}$. In P solve the equation $x^2 - 2x - 5 = 0$.

5.1.5. Let $\mathbb{F}_7 = \{0,1,2,3,4,5,6\}$ be a field. Fill out the multiplication and addition tables of its elements:

+	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2					
2	2						
3	3						
4	4				1		
5	5						
6	6						

,

×	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2			1		
3	0	3					
4	0	4					
5	0	5					
6	0	6					

.

5.1.6. Let F be a field and $f : F \longrightarrow F$ an isomorphism. Prove that the subset $K = \{x \in F \mid f(x) = x\}$ is a subfield of F.

5.1.7. Let $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ be a field. In this field solve the equation $17x = 3$.

5.1.8. Let $P = \{x + y\sqrt{5} \mid x, y \in \mathbb{Q}\}$. Prove that P is a subfield of $\mathbb{R}$. Prove that the mapping $f : P \longrightarrow P$, defined by the rule $f(x + y\sqrt{5}) = x - y\sqrt{5}$ is an isomorphism.

5.1.9. Prove that the fields of the Problems 5.1 and 5.2 are not isomorphic.

5.1.10. Let F be a field of characteristic $p > 0$. Prove that if n is a natural number then $(x + y)^{p^n} = x^{p^n} + y^{p^n}$ for all $x, y \in F$.

5.1.11. Let F be a field and let $f(X) = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n \in F[X]$. Define formal differentiation in $F[X]$ by

$$Df(X) = a_1 + 2a_2X + 3a_3X^2 + \cdots + na_nX^{n-1}.$$

Prove that if $g(X) \in F[X]$ also, then formal differentiation satisfies $D((f + g)(X)) = Df(X) + Dg(X)$.

5.1.12. With the notation and terminology of Problem 5.1.11 prove also that $D((f \cdot g)(X)) = Df(X) \cdot g(X) + f(X) \cdot Dg(X)$.

5.1.13. Let $\mathbb{Q} \leq F \leq \mathbb{C}$ be subfields of $\mathbb{C}$. Prove that if $f : F \longrightarrow F$ is an isomorphism then $f(n/m) = n/m$ for all $n, m \in \mathbb{Z}$ such that $m \neq 0$. (Thus the field $\mathbb{Q}$ is fixed by f and we call f a $\mathbb{Q}$-automorphism of the field F.)

5.1.14. Let F be a field of characteristic $p > 0$ and let $E \cong \mathbb{F}_p$ be the prime subfield of F. Prove that if $\alpha : F \longrightarrow F$ is an isomorphism then $\alpha(u) = u$ for all elements $u \in E$. (Thus α is called an E-automorphism of F.)

5.1.15. Construct a field with 25 elements.

5.1.16. Construct a field with 27 elements.

5.1.17. If p is a prime then how do we go about constructing a field with p^n elements?

5.1.18. Let F be a subfield of the field K. A ring homomorphism $\alpha : K \longrightarrow K$ is called an F-automorphism if α is an injective, surjective homomorphism from K to itself with the property that $\alpha(f) = f$ for all $f \in F$. Prove that the set of all F-automorphisms of K is a group under composition of mappings. (This group is called the Galois group of K over F.)

5.1.19. Compute the Galois group of $\mathbb{C}$ over $\mathbb{R}$.

5.1.20. Prove that if $P = \{a + b\sqrt{2} | a, b \in \mathbb{Q}\}$ then there are exactly 2 $\mathbb{Q}$-automorphisms of P.

5.2 SOME FIELD EXTENSIONS

In Section 5.1, we took the first steps in learning fields, by describing the structure of the prime subfields. A next natural step in the study of fields is to study extensions of fields which leads naturally to the discussion of extensions of prime fields.

Let F be a subfield of a field P and let M be a subset of P. Let $\mathfrak{M}$ denote the family of subfields of P which contain both F and M. We define $F(M) = \cap\mathfrak{M} = \bigcap\{K : K \in \mathfrak{M}\}$. By Corollary 5.1.3, $F(M)$ is a subfield of P. Every subfield of P, containing F and M clearly contains $F(M)$, so that in this sense $F(M)$ is the smallest subfield containing F and M.

Definition 5.2.1. *Let F be a subfield of a field P and let M be a subset of P. The subfield $F(M)$ is called an extension of F, obtained by adjoining the subset M to the field F.*

The first case to consider using this approach is the case when the subset M consists of only one element $\alpha \in P$. In this case, we abbreviate $F(\{\alpha\})$ to the shorter notation $F(\alpha)$ and will talk about the extension of F, obtained by adjoining the element α.

Definition 5.2.2. *An extension obtained by adjoining a single element is called simple extension of F. More precisely, if the element α is algebraic over F, then $F(\alpha)$ is called a simple algebraic extension of F; if α is transcendental over F, then $F(\alpha)$ is called a simple transcendental extension of F.*

The field $\mathbb{Q}(\sqrt{r})$ discussed in Section 5.1 is an example of a simple algebraic extension of the field $\mathbb{Q}$. The field $\mathbb{C}$ of complex numbers is also a simple algebraic extension of $\mathbb{R}$. In fact, $\mathbb{C}$ is the extension obtained by adjoining the root, i, of the polynomial $X^2 + 1$ to the field $\mathbb{R}$.

We next discuss the structure of simple extensions. Let F be a subfield of a field P and let α be an element of P. Consider the evaluation homomorphism $\Phi : F[X] \longrightarrow P$ defined by $\Phi(f(X)) = f(\alpha)$ for each polynomial $f(X) \in F[X]$. In Section 4.4 we proved that the mapping Φ is indeed a homomorphism. By Proposition 4.4.3 we deduce that

$$\mathbf{Im}(\Phi) = \{f(\alpha) \mid f(X) \in F[X]\} = F[\alpha]$$

is a subring of P and

$$\mathbf{Ker}(\Phi) = \{f(X) \mid f(\alpha) = 0_F\} = \mathbf{Ann}_{F[X]}(\alpha)$$

is an ideal of the ring $F[X]$. Furthermore, Theorem 4.4.6 shows that

$$F[\alpha] \cong F[X]/\mathbf{Ann}_{F[X]}(\alpha).$$

More can be said here. Suppose first that α is algebraic over F. Then there exists a polynomial $h(X) \in F[X]$ such that $h(\alpha) = 0_F$. By Theorem 4.2.12, $h(X) = p_1(X)\dots p_m(X)$ where $p_1(X),\dots,p_m(X)$ are irreducible polynomials over F. We have $0_F = h(\alpha) = p_1(\alpha)\dots p_m(\alpha)$. Since P has no zero-divisors, there exists an index j such that $p_j(\alpha) = 0_F$. Let $q(X)$ be another polynomial in $F[X]$ such that $q(\alpha) = 0_F$ and suppose that $q(X) \notin p_j(X)F[X]$. It follows that the ideal $p_j(X)F[X] + q(X)F[X]$ does not coincide with $p_j(X)F[X]$. By Proposition 4.3.17, the ideal $p_j(X)F[X]$ is maximal in the ring $F[X]$ and it follows that $F[X] = p_j(X)F[X] + q(X)F[X]$. We deduce that there exist polynomials $u(X), v(X) \in F[X]$ such that $1_F = u(X)p_j(X) + v(X)q(X)$. From this equation we deduce that

$$1_F = u(\alpha)p_j(\alpha) + v(\alpha)q(\alpha) = u(\alpha)0_F + v(\alpha)0_F = 0_F.$$

This contradiction shows that $q(X) \in p_j(X)F[X]$, which proves that $\mathbf{Ann}_{F[X]}(\alpha) \leq p_j(X)F[X]$. On the other hand, $p_j(X) \in \mathbf{Ann}_{F[X]}(\alpha)$, and since $\mathbf{Ann}_{F[X]}(\alpha)$ is an ideal, $p_j(X)F[X] \leq \mathbf{Ann}_{F[X]}(\alpha)$. Thus $\mathbf{Ann}_{F[X]}(\alpha) = p_j(X)F[X]$. In particular, if $q(X)$ is an irreducible polynomial over F such that $q(\alpha) = 0_F$, then $q(X) = cp_j(X)$ for some element $c \in F$. We let $\mathbf{Irr}(\alpha, X) = a^{-1}p_j(X)$, where a is the leading coefficient of the polynomial $p_j(X)$. Then $\mathbf{Irr}(\alpha, X)$ is the only irreducible polynomial over F whose leading coefficient is 1_F, having α as a root.

We call $\mathbf{Irr}(\alpha, X)$ the *minimal polynomial* of α over F.

The equation $\mathbf{Ann}_{F[X]}(\alpha) = p_j(X)F[X]$ implies that $\mathbf{Ann}_{F[X]}(\alpha) = \mathbf{Irr}(\alpha, X)$ $F[X]$. Then $F[X]/\mathbf{Ann}_{F[X]}(\alpha) \cong F[\alpha]$ is a field, by Propositions 4.3.16 and 4.3.17. Since $1_P = 1_F \in F[\alpha]$, $F[\alpha]$ is a subfield of P. Hence $F(\alpha) = F[\alpha]$ in this case, and this allows us to determine the elements of $F(\alpha)$ quite easily. In fact, if c is a nonzero element of $F(\alpha)$, then $c = f(\alpha)$ for some polynomial $f(X) \in F[X]$. Furthermore, by Theorem 4.2.3, there are polynomials $g(X)$ and $r(X)$ such that $f(X) = g(X)\mathbf{Irr}(\alpha, X) + r(X)$ where either $r(X) = 0_F$ or $\mathbf{deg}\, r(X) < n = \mathbf{deg}\,(\mathbf{Irr}(\alpha, X))$. We have

$$c = f(\alpha) = g(\alpha)\mathbf{Irr}(\alpha, \alpha) + r(\alpha) = g(\alpha)0_F + r(\alpha) = r(\alpha).$$

Thus c can be written as a sum of positive integer powers of α, with coefficients in F, (so we might think of this as a "polynomial" in α), the highest power of which is at most one less than the degree of $\mathbf{Irr}(\alpha,X)$, the minimal polynomial of α over F. Assume that $c = h(\alpha)$ for another polynomial $h(X) \in F[X]$ such that $\mathbf{deg}\, h(X) < n$ and let $s(X) = h(X) - r(X) \in F[X]$. Then either $s(X) = 0_F$ or $\mathbf{deg}\, s(X) < n$. In the first case, we have $h(X) = r(X)$. Suppose that $h(X) \neq r(X)$, so $s(X) \neq 0_F$. We have $s(\alpha) = h(\alpha) - r(\alpha) = 0_F$, which implies that $s(X) \in \mathbf{Ann}_{F[X]}(\alpha)$. The equation $\mathbf{Ann}_{F[X]}(\alpha) = \mathbf{Irr}(\alpha,X)F[X]$ implies that $s(X) = \mathbf{Irr}(\alpha,X)w(X)$ for some polynomial $w(X) \in F[X]$. Since $s(X) \neq 0_F, w(X) \neq 0_F$. Corollary 4.2.2 shows that in this case $\mathbf{deg}\, s(X) \geq n$, and we obtain a contradiction. Hence $h(X) = r(X)$ so the representation of every nonzero element of $F(\alpha)$ in the form $c = r(\alpha)$, where $\mathbf{deg}\, r(X) < \mathbf{deg}\,(\mathbf{Irr}(\alpha,X))$, is unique. We have therefore obtained the following description of a simple algebraic extension.

Theorem 5.2.3. *Let F be a subfield of a field P and let α be an element of P that is algebraic over F. Then $F(\alpha)$ is isomorphic to $F[X]/\mathbf{Irr}(\alpha,X)$. Every element of $F(\alpha)$ has the form $a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1}$ where $a_0, a_1, \ldots, a_{n-1} \in F$, $n = \mathbf{deg}\,(\mathbf{Irr}(\alpha,X))$, and this representation is unique.*

As an example we will build an extension $\mathbb{Q}(\kappa)$ of $\mathbb{Q}$, where $\kappa = \cos(2\pi/5) + i\sin(2\pi/5)$ is a primitive root of unity of degree 5. Since κ satisfies the polynomial equation $X^5 - 1 = 0$ we can easily deduce that the minimal polynomial of κ over $\mathbb{Q}$ is $\mathbf{Irr}(\kappa,X) = X^4 + X^3 + X^2 + X + 1$. (One way to do this is to observe that the nontrivial roots of $X^5 - 1 = 0$ are $\kappa, \kappa^2, \kappa^3, \kappa^4$, none of which are rational. Furthermore the roots occur in complex conjugate pairs and it can be shown that $(X - \kappa)(X - \kappa^4), (X - \kappa^2)(X - \kappa^3) \notin \mathbb{Q}[X]$.) By Theorem 5.2.3 every element of $\mathbb{Q}(\kappa)$ can be uniquely written in the form $a_0 + a_1\kappa + a_2\kappa^2 + a_3\kappa^3$ where $a_0, a_1, a_2, a_3 \in \mathbb{Q}$. Addition here is defined by adding coefficients of the same power as follows.

$$\begin{aligned}&(a_0 + a_1\kappa + a_2\kappa^2 + a_3\kappa^3) + (b_0 + b_1\kappa + b_2\kappa^2 + b_3\kappa^3)\\ &= (a_0 + b_0) + (a_1 + b_1)\kappa + (a_2 + b_2)\kappa^2 + (a_3 + b_3)\kappa^3.\end{aligned}$$

Multiplication is defined via the usual distributive law, but in order to write the product in the form of a sum of powers of κ, where the exponent is at most 3 we need to recall that $\kappa^5 = 1$, so $\kappa^6 = \kappa$ and $\kappa^4 = -1 - \kappa - \kappa^2 - \kappa^3$. Thus we obtain:

$$\begin{aligned}&(a_0 + a_1\kappa + a_2\kappa^2 + a_3\kappa^3)(b_0 + b_1\kappa + b_2\kappa^2 + b_3\kappa^3)\\ &\quad = a_0b_0 + (a_1b_0 + a_0b_1)\kappa + (a_0b_2 + a_1b_1 + a_2b_0)\kappa^2\\ &\qquad + (a_0b_3 + a_1b_2 + a_2b_1 + a_3b_0)\kappa^3 + (a_1b_3 + a_2b_2 + a_3b_1)\kappa^4\\ &\qquad + (a_2b_3 + a_3b_2)\kappa^5 + a_3b_3\kappa^6\end{aligned}$$

We can reduce this further by employing the identities mentioned here. We just illustrate this by the following example.

$$\begin{aligned}(2+\kappa+\kappa^2)(4+2\kappa^3)&\\ &= 8+4\kappa+4\kappa^2+4\kappa^3+2\kappa^4+2\kappa^5\\ &= 8+4\kappa+4\kappa^2+4\kappa^3+2(-1-\kappa-\kappa^2-\kappa^3)+2\\ &= 8+2\kappa+2\kappa^2+2\kappa^3.\end{aligned}$$

We now consider the case when $\alpha \in P$ is transcendental over F. In this case, $\mathbf{Ker}(\Phi) = \mathbf{Ann}_{F[X]}(\alpha) = \{0_F\}$, and Theorem 4.4.4 shows that the mapping Φ is a monomorphism. Thus $F[\alpha] \cong F[X]$ in this case. We let $F\{X\}$ denote the field of fractions of $F[X]$. Now we proceed in the same way as we did in the proof of Theorem 5.1.6 by extending Φ to the mapping $\Phi_1 : F\{X\} \longrightarrow P$ in the following way. Let $f(X)/g(X)$ be an arbitrary element of $F\{X\}$. Then $g(X) \neq 0_F$ and hence $\Phi(g(X)) = g(\alpha) \neq 0_F$. Since P is a field, its nonzero element $g(\alpha)$ is invertible in P and we let

$$\Phi_1(f(X)/g(X)) = f(\alpha)(g(\alpha))^{-1}.$$

Using the arguments of the proof of Theorem 5.1.6 we see that Φ_1 is a well-defined homomorphism. Furthermore, Φ_1 is an extension of Φ, so it is nonzero and Theorem 5.1.5 shows that Φ_1 is a monomorphism. Hence $F\{X\} \cong \mathbf{Im}(\Phi_1)$. Since Φ_1 is an extension of Φ, $\mathbf{Im}(\Phi_1)$ is a subfield of P containing $F[\alpha]$. In particular, $\mathbf{Im}(\Phi_1)$ contains F and the element α, so that $\mathbf{Im}(\Phi_1)$ contains $F(\alpha)$. On the other hand, every subfield of P containing F and α (in particular, $F(\alpha)$), contains the elements $f(\alpha)$ for each polynomial $f(X) \in F[X]$. Hence such a subfield contains all products $f(\alpha)(g(\alpha))^{-1}$, where $f(X), g(X) \in F[X]$ and $g(X) \neq 0_F$. This shows that $\mathbf{Im}(\Phi_1) = F(\alpha)$ and we obtain the following description of a simple transcendental extension.

Theorem 5.2.4. *Let F be a subfield of the field P and suppose that α is an element of P, transcendental over F. Then $F(\alpha)$ is isomorphic to the field $F\{X\}$ of rational functions over F.*

We have proved the previous two theorems on the assumption that there is a field P, containing F as a subfield and containing the element α. But quite often we have to deal with situations where the existence of such a field P must be proved. In other words, the question concerning the existence of the extension $F(\alpha)$ arises naturally here. The answer to this can be obtained with the following version of Theorem 4.4.9.

Proposition 5.2.5. *Let F, K be fields and let $f : F \longrightarrow K$ be a monomorphism. Then there exists a field S such that S is isomorphic to K and F is a subfield of S.*

Proof. We use the construction developed in the proof of Theorem 4.4.9. Since f is a monomorphism, Theorem 5.1.5 shows that $f(1_F) = 1_K$. By Theorem 4.4.9, 1_F is the identity element of S. Theorem 5.1.2 shows that it is sufficient to check that F satisfies (**SF 4**). Suppose that x is a nonzero element of F. Then it has a multiplicative inverse y in F, so $xy = 1_F$. By Theorem 4.4.9 $xy = 1_F = 1_S$, so that y is a multiplicative inverse for x in S. Since $y \in F$, F satisfies (**SF 4**), which proves the result.

Now, the very description of simple algebraic and transcendental extensions obtained in Theorems 5.2.3 and 5.2.4 tells us how to prove their existence.

Theorem 5.2.6. *Let F be a field and let $q(X)$ be a polynomial in $F[X]$. If $q(X)$ is irreducible over F, then there exists an extension S of F, containing a root of $q(X)$.*

Proof. By Proposition 4.3.17, the ideal $M = q(X)F[X]$ is maximal in the polynomial ring $F[X]$. Proposition 4.3.16 shows that the quotient ring $K = F[X]/M$ is a field. Let σ_M denote the natural homomorphism of $F[X]$ onto $F[X]/M$, so that σ_M is defined by $\sigma_M(f(X)) = f(X) + M$ for every polynomial $f(X)$. Let ϕ be the restriction of σ_M to F. Then ϕ is a homomorphism from the field F into the field $F[X]/M$. This homomorphism is nonzero. Indeed, since $q(X)$ is irreducible over F, $\mathbf{deg}\, q(X) \geq 1$. It follows that $a \notin q(X)F[X]$ for every nonzero element a of F and hence $\phi(a) = a + M \neq M$. By Theorem 5.1.5, ϕ is a monomorphism from F into K. An application of Proposition 5.2.5 shows that there exists a field S such that F is a subfield of S and $S \cong K$. Let $q(X) = v_0 + v_1X + \cdots + v_kX^k$. We now use the notation of the proof of Theorem 4.4.9 and set $\alpha = X + M$. Then

$$\begin{aligned} q(\alpha) &= v_0 + v_1(X+M) + \cdots + v_k(X+M)^k \\ &= (v_0 + v_1X + \cdots + v_kX^k) + M = q(X) + M = M. \end{aligned}$$

The ideal M is the zero element in the quotient ring $F[X]/M$, and therefore its preimage in S is $0_S = 0_F$. So in the field S we obtain the equation $q(\alpha) = 0_F$.

With the existence of a transcendental extension the situation is even simpler. Consider the field $P = F\{X\}$ of all rational functions over F. The field F is naturally embedded in the polynomial ring $F[X]$ and hence in $F\{X\}$. In the

field P the element X is not the root of any polynomial with coefficients in F. Thus P is a simple transcendental extension of F.

Once theorems concerning existence of extensions have been established it is natural to ask about the number of nonisomorphic extensions. However, in Theorem 5.2.3 we show that every simple algebraic extension $F(\alpha)$ is isomorphic to $F[X]/\mathbf{Irr}(\alpha, X)$ and hence all such extensions are isomorphic. In Theorem 5.2.4 we show that every simple transcendental extension $F(\alpha)$ is isomorphic to $F\{X\}$, and hence all such extensions are also isomorphic.

Exercise Set 5.2

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

5.2.1. Prove that $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3})$.

5.2.2. Prove that $\mathbb{Q}(\sqrt{5}, \sqrt{7}) = \mathbb{Q}(\sqrt{5} + \sqrt{7})$.

5.2.3. Describe the elements of the field $\mathbb{Q}(\sqrt{7})$.

5.2.4. Describe the elements of the field $\mathbb{Q}(\sqrt[3]{5})$.

5.2.5. Describe the element of the field $\mathbb{Q}(\pi)$.

5.2.6. Let $\alpha \in \mathbb{C}$ be algebraic over $\mathbb{Q}$ and let $f(X)$ be the minimal polynomial of α over $\mathbb{Q}$. Let $\Phi : \mathbb{C} \longrightarrow \mathbb{C}$ be a $\mathbb{Q}$-automorphism of $\mathbb{C}$. Prove that $\Phi(\alpha)$ is another root of the polynomial $f(X)$.

5.2.7. Prove that if $\theta : \mathbb{Q}(\sqrt[3]{2}) \longrightarrow : \mathbb{Q}(\sqrt[3]{2})$ is an isomorphism then θ is the identity map.

5.2.8. Let $\kappa \neq 1$ be a primitive 5th root of unity, so $\kappa^5 = 1$. Prove that for the field $\mathbb{Q}(\kappa)$ there are exactly four $\mathbb{Q}$-automorphisms and find them.

5.2.9. Find the minimal polynomial of $\sqrt{2} + \sqrt{3}$.

5.2.10. Prove that $\mathbb{Q}(\sqrt{5})$ has exactly two $\mathbb{Q}$-automorphisms.

5.3 FIELDS OF ALGEBRAIC NUMBERS

In this section we continue our work on the theory of fields by extending the approach outlined in Section 5.2. In Section 5.1 we described the structure of the prime subfield of a field and every field can be considered as the result of adjoining a set of elements to the prime field. In Section 5.2 we made the first natural step in this direction by studying extensions obtained by adjoining

one element and this allowed us to obtain a description of the structure of simple extensions. The natural next step is to consider extensions obtained by adjoining a finite number of elements to a given field. The situation here is much more diverse. We can show here only a very small part of the big picture, collecting only the most essential concepts and results. We note also that the source of all of this work, historically, arose from the problem of finding a formula linking the roots of a polynomial with its coefficients. This classical work, and its many consequences, was carried out by many famous mathematicians, among them Galois, who laid the foundations of the modern theory of fields, and after whom an extensive and vibrant theory has been named.

Let F be a subfield of a field P. Then P is an abelian group under addition. We define an action of F on P as follows. For every $\alpha \in F, a \in P$ we define a mapping from $F \times P$ to P by $(\alpha, a) \longmapsto \alpha a$ (the product of these elements in the field P). By the distributive law in P we have $(\alpha+\beta)a = \alpha a + \beta a$ and $\alpha(a+b) = \alpha a + \alpha b$, for all $\alpha, \beta \in F, a, b \in P$. Likewise the associative law gives us $\alpha(\beta a) = (\alpha\beta)a$, and since $1_F = 1_P$, we have $1_F a = a$. This shows that P satisfies the axioms for a vector space and hence we can think of P also as a vector space over F. This creates a lot of flexibility in the way we can think about P. At times we may think of P as a field, but also sometimes as a vector space over F.

Definition 5.3.1. *Let F be a subfield of a field P. We say that P is a finite (field) extension of F if P is a finite dimensional vector space over F. The dimension* $\mathbf{dim}_F(P)$ *is called the degree of the extension P over F and will be denoted by* $[P : F]$.

Suppose that P is a finite field extension of F, and let $M = \{a_1, \ldots, a_n\}$ be a basis for P as a vector space over F. Then clearly $P = F(M)$, since every element of P is now an F-linear combination of $a_1, a_2, \ldots, a_n$.

Definition 5.3.2. *Let F be a subfield of a field P. We say that P is finitely generated over F, or P is a finitely generated extension of F, if there exists a finite subset M of P such that $P = F(M)$.*

Thus we can see that every finite extension of a field F is finitely generated over F.

Suppose now that P is a simple algebraic extension of F, say $P = F(\gamma)$ where γ is an algebraic element of P over F. Let $\mathbf{Irr}(\gamma, X) = c_0 + c_1 X + \cdots + c_n X^n$ where $c_0, c_1, \ldots, c_n \in F$. By Theorem 5.2.3, every element u of P can be written uniquely in the form $u = a_0 + a_1\gamma + \cdots + a_{n-1}\gamma^{n-1}$ where $a_0, a_1, \ldots, a_{n-1} \in F$. This shows that the elements $1_F, \gamma, \ldots, \gamma^{n-1}$ form a basis

for the vector space P over F. Hence every simple algebraic extension is a finite extension.

We need some properties of finite extensions.

Theorem 5.3.3. *Let F be a subfield of a field P and let P be a subfield of a field K. If K is a finite field extension of F, then K is a finite field extension of P, and P is a finite field extension of F. Conversely, suppose that K is a finite field extension of P and P is a finite field extension of F. Then K is a finite field extension of F. Moreover,* $[K : F] = [K : P][P : F]$.

Proof. Assume that K is a finite field extension of F and let $\{a_1, \ldots, a_n\}$ be a basis for K over F. Clearly P is a subspace of K and therefore

$$[P : F] = \mathbf{dim}_F P \leq \mathbf{dim}_F K = [K : F].$$

Every element of K has the form $\lambda_1 a_1 + \ldots \lambda_n a_n$ where $\lambda_1, \ldots, \lambda_n \in F$. It follows that every element of K is a linear combination of the elements $a_1, \ldots, a_n$ with coefficients from the subfield F, so these coefficients also lie in P. Thus, the vector space K is generated by the elements $a_1, \ldots, a_n$ when we think of K as a vector space over P. Since finitely generated vector spaces are finite dimensional, it follows that $\mathbf{dim}_P K = [K : P]$ is finite.

Conversely, let $\{c_1, \ldots, c_k\}$ be a basis of K over P and $\{b_1, \ldots, b_t\}$ be a basis of P over F. If a is an arbitrary element of K then

$$a = \sum_{1 \leq j \leq k} \lambda_j c_j \text{ where } \lambda_j \in P, \text{ for } 1 \leq j \leq k.$$

Further, for each j we have

$$\lambda_j = \sum_{1 \leq m \leq t} \mu_{jm} b_m, \text{ where } \mu_{jm} \in F, \text{ for } 1 \leq m \leq t, 1 \leq j \leq k.$$

It follows that

$$a = \sum_{1 \leq j \leq k} \lambda_j c_j = \sum_{1 \leq j \leq k} \left(\sum_{1 \leq m \leq t} \mu_{jm} b_m \right) c_j = \sum_{1 \leq j \leq k} \sum_{1 \leq m \leq t} \mu_{jm} (b_m c_j).$$

This shows that the elements $b_m c_j$, for $1 \leq m \leq t, 1 \leq j \leq k$, span the F-vector space K. Since K is finitely generated it is finite dimensional over F. We show that the subset $\mathcal{M} = \{b_m c_j \mid 1 \leq m \leq t, 1 \leq j \leq k\}$ is a basis for K over F. We have already seen that this set of elements spans the vector space K. We now show that this set of elements is linearly independent. Indeed, for $1 \leq m \leq t, 1 \leq j \leq$

k, let ν_{jm} be elements of the field F such that $\sum_{1\le j\le k}\sum_{1\le m\le t}\nu_{jm}(b_m c_j) = 0_F$. We have

$$0_F = \sum_{1\le j\le k}\sum_{1\le m\le t}\nu_{jm}(b_m c_j) = \sum_{1\le j\le k}\left(\sum_{1\le m\le t}\nu_{jm}b_m\right)c_j.$$

Since $\{c_1,\ldots,c_k\}$ is a basis for K over P, and $\sum_{1\le m\le t}\nu_{jm}(b_m)\in P$ we deduce that $\sum_{1\le m\le t}\nu_{jm}b_m = 0_F$, for each j such that $1\le j\le k$. Likewise the subset $\{b_1,\ldots,b_t\}$ is a basis of P over F, and therefore the last equation implies that $\nu_{jm} = 0_F$ whenever $1\le m\le t$, $1\le j\le k$. It follows that the elements $\{b_m c_j \mid 1\le m\le t, 1\le j\le k\}$ are linearly independent over F and hence $\mathcal{M}$ is a basis of K over F. Thus

$$[K:F] = \mathbf{dim}_F K = kt = \mathbf{dim}_P K\cdot\mathbf{dim}_F P = [K:P][P:F].$$

This completes the proof.

As we can see in the proof of Theorem 5.3.3, once we have a basis for K over P and a basis for P over F, it is actually very easy to find a basis for K over F—you just multiply the corresponding basis elements together.

Definition 5.3.4. *Let F be a subfield of a field P. We say that P is an algebraic extension of F, if every element of P is algebraic over F.*

Corollary 5.3.5. *Let P be a finite field extension of the field F. Then P is an algebraic extension of F.*

Proof. Assume that $[P:F] = n$ and let γ be an arbitrary element of P. Consider the elements $1_F = \gamma^0, \gamma, \ldots, \gamma^n$. Since $\mathbf{dim}_F P = n$, the subset of $n+1$ elements $\{1_F, , \gamma, \ldots, \gamma^n\}$ is linearly dependent over F. Hence there exist elements $a_0,\ldots,a_n\in F$, at least one of which is nonzero, such that $a_0 + a_1\gamma + \cdots + a_n\gamma^n = 0_F$. Let $g(X) = a_0 + a_1X + \cdots + a_nX^n$. Then $g(X)$ is a nonzero polynomial in $F[X]$ and γ is a root of this polynomial. Hence γ is algebraic over F for each element of P.

It is a very easy exercise to deduce the following result.

Corollary 5.3.6. *Let P be a finite field extension of the field F. Then $P = F(\gamma_1,\ldots,\gamma_n)$, for certain elements $\gamma_1,\ldots,\gamma_n$ of P and the elements $\gamma_1,\ldots,\gamma_n$ are algebraic over F.*

The converse is also true, so that we can deduce the following characterization of finite field extensions.

Theorem 5.3.7. *Let F be a subfield of a field P. Then P is a finite extension of F if and only if $P = F(\gamma_1, \ldots, \gamma_n)$ and all elements $\gamma_1, \ldots, \gamma_n$ are algebraic over F.*

Proof. If P is finite extension of F, then the result follows from Corollary 5.3.6.

Conversely, let $P = F(\gamma_1, \ldots, \gamma_n)$ where the elements $\gamma_1, \ldots \gamma_n$ are algebraic over F. For each i such that $1 \leq i \leq n$, let $P_i = F(\gamma_1, \ldots, \gamma_i)$ and consider the following chain of extensions

$$F = P_0 \leq P_1 \leq P_2 \leq \cdots \leq P_{n-1} \leq F(\gamma_1, \ldots, \gamma_n) = P.$$

Notice that $P_{i+1} = P_i(\gamma_{i+1})$ for each i such that $0 \leq i \leq n-1$ and, as we noted earlier, every simple algebraic extension is a finite extension. Thus P_1 is a finite extension of F. We have $P_2 = F(\gamma_1, \gamma_2) = P_1(\gamma_2)$. The element γ_2 is a root of some polynomial $f_2(X) \in F[X]$. Since $F \leq P_1, f_2(X) \in P_1[X]$, so that γ_2 is algebraic over P_1. It follows that $P_2 = P_1(\gamma_2)$ is a finite extension of P_1. An application of Theorem 5.3.3 shows that $[P_2 : P_0] = [P_2 : P_1] \cdot [P_1 : P_0]$ so P_2 is a finite extension of F and an inductive argument, based on similar ideas allows us to deduce that P_{i+1} is a finite extension of P_i for $0 \leq i \leq n-1$. It follows, using Theorem 5.3.3, that P is a finite field extension of F.

This theorem implies the following important corollaries.

Corollary 5.3.8. *Let F be a subfield of a field P and let P be a subfield of a field K. If K is algebraic over P and P is algebraic over F, then K is an algebraic extension of F.*

Proof. Let α be an arbitrary element of K. Then there exists a polynomial $f(X) \in P[X]$ such that $f(\alpha) = 0_F$. Let $f(X) = a_0 + a_1X + \cdots + a_nX^n$. Since $a_j \in P$ it follows that a_j is algebraic over F, for $0 \leq j \leq n$. By Theorem 5.3.7, $S = F(a_0, a_1, \ldots, a_n)$ is a finite field extension of F. Then α is algebraic over S, and hence $S(\alpha)$ is a finite field extension of S. Theorem 5.3.3 shows that $[S(\alpha) : F] = [S(\alpha) : S] \cdot [S : F]$ so $S(\alpha)$ is a finite field extension of F and Corollary 5.3.5 proves that α is algebraic over F.

Corollary 5.3.9. *Let F be a subfield of a field P. Then*

$$\mathbf{Alg}_F(P) = \{\alpha \mid \alpha \in P \text{ and } \alpha \text{ is algebraic over } F\}$$

is a subfield of P.

Proof. Clearly $F \leq \mathbf{Alg}_F(P)$. In particular, $1_F = 1_P \in \mathbf{Alg}_F(P)$, so that $\mathbf{Alg}_F(P)$ satisfies (**SF 3**). Let $\alpha, \beta \in \mathbf{Alg}_F(P)$ and put $K = F(\alpha, \beta)$. By Theorem 5.3.7, K is a finite field extension of F. Corollary 5.3.5 implies that K is an algebraic extension of F. It follows that $\alpha - \beta, \alpha\beta$, which belong to K, are algebraic over F. Thus $\alpha - \beta, \alpha\beta \in \mathbf{Alg}_F(P)$, and $\mathbf{Alg}_F(P)$ satisfies (**SF 1**) and (**SF 2**). Finally, note that if α is nonzero then, since K is a subfield, $\alpha^{-1} \in K$. The inclusion $K \leq \mathbf{Alg}_F(P)$ shows that $\alpha^{-1} \in \mathbf{Alg}_F(P)$, so that $\mathbf{Alg}_F(P)$ satisfies the condition (**SF 4**). Now Theorem 5.1.2 proves the result.

Definition 5.3.10. *Let F be a subfield of a field P. The subfield $\mathbf{Alg}_F(P)$ is called the algebraic closure of F in P.*

The algebraic closure of $\mathbb{Q}$ in $\mathbb{C}$ is called the field of algebraic numbers. There is some ambiguity in the terminology here. Quite often, in common usage, a field of algebraic numbers means a finite extension of the field of rational numbers. *The* field of algebraic numbers refers to the set of all complex numbers which are algebraic over $\mathbb{Q}$. Algebraic numbers have been studied for a long time and the topic represents a very well-developed area of algebra and number theory. A vast literature has been dedicated to algebraic numbers. Therefore, we will not delve into this topic, but present only some of the important properties of algebraic numbers.

Corollary 5.3.11. *The field of all algebraic numbers is algebraically closed.*

Proof. Let $f(X) = a_0 + a_1X + \cdots + a_nX^n$ be a polynomial whose coefficients $a_0, a_1, \ldots, a_n$ belong to $\mathbf{Alg}_{\mathbb{Q}}(\mathbb{C})$. By Theorem 4.2.10, every root α of this polynomial is a complex number. Since $\mathbf{Alg}_{\mathbb{Q}}(\mathbb{C})(\alpha)$ is algebraic over $\mathbf{Alg}_{\mathbb{Q}}(\mathbb{C})$ and $\mathbf{Alg}_{\mathbb{Q}}(\mathbb{C})$ is algebraic over $\mathbb{Q}$, Corollary 5.3.8 shows that α is algebraic over $\mathbb{Q}$. Thus $\alpha \in \mathbf{Alg}_{\mathbb{Q}}(\mathbb{C})$, which proves the result.

Theorem 5.3.12. *The field of all algebraic numbers is countable.*

Proof. Let α be an algebraic number and let $f(X) = a_0 + a_1X + \cdots + a_nX^n$ be a polynomial with rational coefficients such that $f(\alpha) = 0$. Let k be the product of all the denominators of the elements $a_0, a_1, \ldots, a_n$. Then all coefficients of the polynomial $g(X) = ka_0 + ka_1X + \cdots + ka_nX^n$ are integers, and clearly $g(\alpha) = 0$. Thus, for every algebraic number there exists a polynomial with integer coefficients having this number as a root.

Now consider the set $\mathbb{Z}[X]$ of all polynomials with integer coefficients. For every positive integer n, let L_n denote the subset of $\mathbb{Z}[X]$ consisting of all polynomials $f(X) = c_0 + c_1X + \cdots + c_mX^m$ such that $\mathbf{deg}\, f(X) \leq n$ and $|c_0| + |c_1| + \cdots + |c_m| \leq n$. Since we do not ascribe any power to the zero polynomial

we shall simply assume that $0 \in L_1$, so that $L_1 = \{0, 1, -1, X, -X\}$. Clearly, each of the subsets L_n is finite. For example,

$$\begin{aligned} L_2 = \{&1, -1, 2, -2, X, -X, 1+X, 1-X, -1+X, -1-X, 2X, -2X, \\ &1+X^2, 1-X^2, -1+X^2, -1-X^2, X+X^2, -X+X^2, X-X^2, \\ &-X-X^2, 2X^2, -2X^2\}. \end{aligned}$$

Every polynomial $g(X) \in \mathbb{Z}[X]$ belongs to some subset L_n. Indeed, let $g(X) = b_0 + b_1X + \cdots + b_tX^t$. Without loss of generality we can assume that $b_t \neq 0$, so that $\mathbf{deg}\ g(X) = t$. Let $|b_0| + |b_1| + \cdots + |b_t| = k$. If $k \leq t$, then $g(X) \in L_t$. If $t \leq k$, then $g(X) \in L_k$.

We now begin to enumerate the algebraic numbers by using the set $\mathbb{N}$. First we write out all the algebraic numbers that are roots of polynomials of the subset L_1. The only such number is 0 and we associate this with the number 1. Then we write down the algebraic numbers that are roots of polynomials that form the subset L_2. These numbers are $0, 1, -1, i, -i$. Since 0 has already been numbered, we index the rest like this:

$$0 \longrightarrow 1, 1 \longrightarrow 2, -1 \longrightarrow 3, i \longrightarrow 4, -i \longrightarrow 5.$$

Next, we consider the algebraic numbers that are roots of polynomials that form the subset L_3. We write down the ones that have not yet been considered. These numbers form a finite set, and we can continue to number them using natural numbers greater than 5. Next, we consider those algebraic numbers which are roots of polynomials that form the subset L_4 and write down the ones that have not yet been considered. The set of them is finite, and we can count them. This process continues. As we noted earlier, every algebraic number is a root of a polynomial with integer coefficients, and as each such polynomial is contained in a subset of L_n, then at some stage every algebraic number will be counted so it will get a natural number as an index. This proves the result.

We note the following important property of algebraic numbers.

Theorem 5.3.13. *Let $\gamma_1, \ldots, \gamma_n$ be algebraic numbers and let $F = \mathbb{Q}(\gamma_1, \ldots, \gamma_n)$. Then F contains an element λ such that $F = \mathbb{Q}(\lambda)$.*

Proof. We will use induction on n, the case $n = 1$ being clear, since we can set $\lambda = \gamma_1$. Suppose that $n = 2$. Let $f_1(X)$ (respectively $f_2(X)$) be the minimal polynomial of γ_1 (respectively γ_2) over $\mathbb{Q}$, and let $\beta_1 = \gamma_1, \beta_2, \ldots, \beta_n$ (respectively $\eta_1 = \gamma_2, \eta_2, \ldots, \eta_m$) be all the roots of $f_1(X)$ (respectively $f_2(X)$). Since $f_1(X)$ is irreducible over $\mathbb{Q}$, all its roots are pairwise distinct. Indeed, if we assume the contrary, then $\mathbf{deg}\, f_1(X) \geq 2$ and $f_1(X)$ and its derivative $f_1'(X)$

have a root in common. Moreover, $f_1'(X) \neq 0$ and $\mathbf{deg}\, f_1'(X) = \mathbf{deg}\, f_1(X) - 1$. It follows that $f_1'(X) \notin f_1(X)\mathbb{Q}[X]$. Since $f_1(X)$ is irreducible over $\mathbb{Q}$, Proposition 4.3.17 shows that the ideal $f_1(X)\mathbb{Q}[X]$ is maximal in $\mathbb{Q}[X]$. The fact that $f_1'(X) \notin f_1(X)\mathbb{Q}[X]$ implies that $f_1(X)\mathbb{Q}[X] + f_1'(X)\mathbb{Q}[X] = \mathbb{Q}[X]$. Then there exist polynomials $g(X), h(X) \in Q[X]$ such that $1 = g(X)f_1(X) + h(X)f_1'(X)$. If β is a common root of $f_1(X)$ and $f_1'(X)$, then we obtain

$$1 = g(\beta)f_1(\beta) + h(\beta)f_1'(\beta) = g(\beta)0 + h(\beta)0 = 0,$$

and we arrive at a contradiction. This contradiction shows that the roots of the polynomial $f_1(X)$ are pairwise distinct. The same is the case for the polynomial $f_2(X)$.

We now proceed to the choice of λ. The number λ we will choose has the form $\lambda = \gamma_1 + c\gamma_2$, where c is a rational number which we now select as follows. We choose the nonzero number c such that all the numbers $\kappa_{jk} = \beta_j + c\eta_k$ are pairwise distinct, for $1 \leq j \leq n, 1 \leq k \leq m$. Such a choice of c is indeed possible. In fact, assume that $\beta_j + c\eta_k = \beta_t + c\eta_s$ for some indices j, k, t, s such that $(j,k) \neq (t,s)$. Note that $j = t$ if and only if $k = s$. Thus $j \neq t$ if and only if $k \neq s$ and then $c = \frac{\beta_t - \beta_j}{\eta_k - \eta_s}$. This equation tells us what the choice of c should be. If we choose c to be a rational number that does not coincide with any one of the numbers $\frac{\beta_t - \beta_j}{\eta_t - \eta_j}$ then the numbers κ_{jk} will be distinct. However the numbers κ_{jk} form a finite set, and the field $\mathbb{Q}$ is infinite, so such a c can be chosen.

Having chosen c, as above, we show that $\mathbb{Q}(\lambda) = \mathbb{Q}(\gamma_1, \gamma_2)$. Since $\lambda = \gamma_1 + c\gamma_2 \in \mathbb{Q}(\gamma_1, \gamma_2)$, it follows that $\mathbb{Q}(\lambda) \leq \mathbb{Q}(\gamma_1, \gamma_2)$. Now consider the polynomial $v(X) = f_1(\lambda - cX)$. Then $v(X)$ is a polynomial, whose coefficients belong to the field $\mathbb{Q}(\lambda)$. We have $v(\gamma_2) = f_1(\gamma_1 + c\gamma_2 - c\gamma_2) = f_1(\gamma_1) = 0$. It follows that the polynomials $f_2(X)$ and $v(X)$ have the common root γ_2. Suppose that some other root η_k of the polynomial $f_2(X)$ is also a root of $v(X)$. Then $0 = v(\eta_k) = f_1(\lambda - c\eta_k) = f_1(\beta_1 + c\eta_1 - c\eta_k)$, since $\beta_1 = \gamma_1$ and $\eta_1 = \gamma_2$, so $\beta_1 + c\eta_1 - c\eta_k$ is a root of $f_1(X)$. Now the roots of the polynomial $f_1(X)$ are $\beta_1, \beta_2, \ldots, \beta_n$. Thus we obtain $\beta_1 + c\eta_1 - c\eta_k = \beta_j$ for some j, such that $1 \leq j \leq n$. But in this case we have $\beta_1 + c\eta_1 = \beta_j + c\eta_k$, which is impossible, by the choice of c. This contradiction shows that γ_2 is the only common root of the polynomials $f_2(X)$ and $v(X)$. By Theorem 4.2.10 and Corollary 4.2.9 we have the following equations over the field $\mathbb{C}$ of complex numbers

$$f_2(X) = (X - \gamma_2)(X - \eta_2)(X - \eta_3)\ldots(X - \eta_m),$$
$$v(X) = (X - \gamma_2)(X - \nu_2)(X - \nu_3)\ldots(X - \nu_n),$$

for certain elements $\nu_2, \ldots, \nu_n \in \mathbb{C}$. As we have just shown, $\eta_j \neq \nu_k$ for all j, k such that $2 \leq j \leq m, 2 \leq k \leq n$ and it follows that $\mathbf{GCD}(f_2(X), v(X)) = (X - \gamma_2)$.

It follows that there are polynomials $a(X)$ and $b(X)$ such that $f_2(X)a(X) + v(X)b(X) = X - \gamma_2$. The greatest common divisor of $f_2(X)$ and $v(X)$ can be found by using the Euclidean algorithm. Recall that the coefficients of $f_2(X)$ and $v(X)$ belong to the field $\mathbb{Q}(\lambda)$. Using the Euclidean algorithm, we can find the coefficients of the polynomials $a(X)$ and $b(X)$ that we need using only the operations of addition, subtraction, multiplication, and division. Therefore these coefficients are elements of the field $\mathbb{Q}(\lambda)$. This means that the coefficients of the polynomial $\mathbf{GCD}(f_2(X), v(X))$ also belong to the field $\mathbb{Q}(\lambda)$. The equation $\mathbf{GCD}(f_2(X), v(X)) = (X - \gamma_2)$ now shows that $\gamma_2 \in \mathbb{Q}(\lambda)$. Then also $\gamma_1 = \lambda - c\gamma_2 \in \mathbb{Q}(\lambda)$. This proves that $\mathbb{Q}(\gamma_1, \gamma_2) = \mathbb{Q}(\lambda)$.

Suppose now that $n > 2$ and we have already proved our result for all sets of algebraic numbers with fewer than n elements. In particular, this assertion is true for the subfield $\mathbb{Q}(\gamma_1, \ldots, \gamma_{n-1})$. By the inductive hypothesis there exists an element λ_1 such that $\mathbb{Q}(\gamma_1, \ldots, \gamma_{n-1}) = \mathbb{Q}(\lambda_1)$. Consider now the subfield $\mathbb{Q}(\lambda_1, \gamma_n)$. By the case $n = 2$, there exists an element λ such that $\mathbb{Q}(\lambda_1, \gamma_n) = \mathbb{Q}(\lambda)$. It now remains only to note that

$$\mathbb{Q}(\gamma_1, \ldots, \gamma_n) = \mathbb{Q}(\gamma_1, \ldots, \gamma_{n-1})(\gamma_n) = \mathbb{Q}(\lambda_1)(\gamma_n) = \mathbb{Q}(\lambda_1, \gamma_n) = \mathbb{Q}(\lambda).$$

As can be seen from the proof, the field $\mathbb{Q}$ can be replaced by any subfield K of the field of complex numbers. Thus, every finite extension of K is a simple algebraic extension.

We have already defined the relative algebraic closure, that is the algebraic closure of one field in an extension field. However, for the field $\mathbb{Q}$ of rationals we have found its absolute algebraic closure, the minimal algebraically closed field containing $\mathbb{Q}$. It is the field of all algebraic numbers. We will generalize this situation now.

Definition 5.3.14. *Let F be a field. A field extension P is called an algebraic closure of F if P is algebraically closed and every proper subfield of P containing F is not algebraically closed.*

In other words, the algebraic closure of F is the minimal algebraically closed field containing F. The following theorem gives the answer to the question of the existence and uniqueness of algebraic closures.

Theorem 5.3.15. *Let F be a field. Then there exists an algebraically closed field containing F, and hence there exists an algebraic closure of F. If P, K are two algebraic closures of F, then there exists an isomorphism $\theta : P \longrightarrow K$ such that $\theta(a) = a$ for each element $a \in F$.*

We omit the proof of this theorem.

Let F be a subfield of $\mathbb{C}$ and $f(X)$ be a polynomial whose coefficients belong to F. By Theorem 4.2.10, all roots $\alpha_1, \ldots, \alpha_n$ of $f(X)$ belong to $\mathbb{C}$. We can form the subfield $F(\alpha_1, \ldots, \alpha_n)$. This is the minimal subfield among those containing F, over which $f(X)$ splits into a product of linear factors. This leads us to another natural generalization.

Definition 5.3.16. *Let F be a field and let $f(X) \in F[X]$ be a polynomial. A field extension P of F is called a splitting field of $f(X)$ over F, if P contains all roots of $f(X)$ and every proper subfield of P containing F, does not contain at least one root of $f(X)$.*

The question of the existence of splitting fields is solved at once, because for a field F there exists an algebraic closure. The question of the uniqueness of a splitting field (to within isomorphism) also has a positive solution. Its proof is a very technically sophisticated version of the proof of Theorem 5.2.6 and so we do not present the proof here.

Exercise Set 5.3

In each of the following questions explain your reasoning, either by giving a proof of your assertion or a counterexample.

5.3.1. Find a basis for $\mathbb{Q}(\sqrt{3})$ over $\mathbb{Q}$ and hence find the dimension of $\mathbb{Q}(\sqrt{3})$ over $\mathbb{Q}$.

5.3.2. Let $f(X) = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$ be a polynomial with integer coefficients. Eisenstein's Criterion asserts that if there is a prime p for which p^2 does not divide a_0, $p|a_i$ for $i < n$ and p does not divide a_n, then $f(X)$ is irreducible over $\mathbb{Q}$. Prove that $f(X) = X^5/2 + 9X^3/2 + 3X^2 - 27X + 3/4$ is irreducible over $\mathbb{Q}$.

5.3.3. Prove using Eisenstein's Criterion that if p is a prime then $f(X) = 1 + X + X^2 + \cdots + X^{p-1}$ is irreducible over $\mathbb{Q}$.

5.3.4. Let $\omega \neq 1$ be a solution of the equation $X^3 = 1$. (Such a solution is called a primitive cube root of unity.) Find a basis for $\mathbb{Q}(\omega)$ over $\mathbb{Q}$ and also find the dimension of $\mathbb{Q}(\omega)$ over $\mathbb{Q}$.

5.3.5. Let $\xi \neq 1$ be a solution of $X^7 = 1$. Find $[\mathbb{Q}(\xi) : \mathbb{Q}]$. Find a basis for $\mathbb{Q}(\xi)$ over $\mathbb{Q}$.

5.3.6. Find $[\mathbb{Q}(\sqrt[4]{7}) : \mathbb{Q}]$ and find a basis for $\mathbb{Q}(\sqrt[4]{7})$ over $\mathbb{Q}$.

5.3.7. Find the degree of the extension $\mathbb{Q}(\sqrt[5]{3}, \sqrt[3]{2})$ over $\mathbb{Q}$ and determine a basis.

5.3.8. Let F be an extension of the field E. Let K be an extension of the field F and let $\alpha \in K$. Then α has a minimal polynomial $m_E(X)$ over E (thus $m_E(X)$ is irreducible in $E[X]$) and also α has a minimal polynomial $m_F(X)$ over F. Prove that $m_F(X)$ divides $m_E(X)$.

5.3.9. Let p be a prime and let κ be a primitive p th root of unity, so $\kappa^p = 1$. Find $[\mathbb{Q}(\kappa) : \mathbb{Q}]$.

5.3.10. Prove that there are exactly four $\mathbb{Q}$-automorphisms of $\mathbb{Q}(\sqrt{2}, \sqrt{5})$ and find them.

5.3.11. Prove that if K is a field extension of F and $[K : F]$ is prime then K is a simple extension of F and that there are no subfields strictly between F and K.

5.3.12. Prove that for every natural number n there is a polynomial with integer coefficients and of degree n that is irreducible over $\mathbb{Q}$.

5.3.13. If F is a finite field then prove that F must have p^n elements, where p is the characteristic of F.

HINTS AND ANSWERS TO SELECTED EXERCISES

CHAPTER 1

1.1.1. Solution. (i) No, not valid for all sets.
(ii) No not valid for arbitrary sets.

1.1.3. Solution. Let $A = \{1\}, B = \{1,2,3\}$. Then $A \subseteq B, A \neq B$. Put $C = \{B\}$. Then $B \in C$. Now put $D = \{C\}$.

1.1.5. Solution. $\mathbb{Z} \backslash A = \mathbb{Z}^- \cup \{x \in \mathbb{Z} | x = 2y+1$ for some $y \geq 0\}$ where $\mathbb{Z}^-$ is the set of all negative integers.
$\mathbb{Z} \backslash (A \cap B) = \mathbb{Z}$.

1.1.7. Proof. For any complex number α we have $f(\alpha) = g(\alpha)h(\alpha)$.

1.1.9. Proof. Let the set C be an element of $\mathfrak{B}(A \cap B)$. Then $C \in \mathfrak{B}(A)$ and $C \in \mathfrak{B}(B)$. Then $C \in \mathfrak{B}(A) \cap \mathfrak{B}(B)$.

1.1.11. Proof. Let $x \in A \setminus (A \setminus B)$. Then $x \in A$ and $x \notin A \setminus B$. Hence $x \in A$. Since $x \notin A \setminus B$ it follows that $x \in B$, because if $x \notin B$ then $x \in A \setminus B$, a contradiction. Therefore $x \in A$ and $x \in B$ so $x \in A \cap B$. Conversely, suppose that $x \in A \cap B$. Then $x \in A$ and $x \in B$. If $x \in A \setminus B$ then $x \in A$

An Introduction to Essential Algebraic Structures, First Edition. Martyn R. Dixon, Leonid A. Kurdachenko and Igor Ya. Subbotin.

and $x \notin B$, a contradiction. Thus $x \in$ and $x \notin A \setminus B$, which means that $x \in A \setminus (A \setminus B)$.

1.1.13. Solution. $\cap_{n\geq 1} A_n = \{0\}$.

1.1.15. Solution. $(A \cap B) \setminus C = A \cap B \neq \emptyset$.

1.1.17. Solution. $(A \cap B) \times C = (A \times C) \cap (B \times C)$ holds.

1.1.19. Solution. Let $x \in A \triangle B$. Then $x \in A \cup B$ and $x \notin A \cap B$. Suppose that $x \in A$. Then $x \notin B$, otherwise $x \in A \cap B$, a contradiction. Thus if $x \in A$ then $x \notin B$ so $x \in A \setminus B$. By a similar argument, if $x \in B$ then $x \notin A$ so $x \in B \setminus A$. Hence $A \triangle B \subseteq (A \cup B) \setminus (A \cap B)$. Conversely, if $x \in (A \setminus B) \cup (B \setminus A)$ then either $x \in A$ and $x \notin B$ or $x \in B$ and $x \notin A$. Thus $x \in A \cup B$ and if $x \in A$ then $x \notin B$ so $x \notin A \cap B$. Likewise if $x \in B$ then $x \notin A$ so $x \notin A \cap B$. In any case $x \in A \triangle B$.

One way to prove the second part is to observe:

$$\begin{aligned} A \cup (B \triangle C) &= A \cap ((B \cup C) \setminus (B \cap C)) \\ &= A \cap ((B \setminus C) \cup (C \setminus B)) \\ &= [A \cap (B \setminus C)] \cup [A \cap (C \setminus B)] \text{ by distributive law} \\ &= [(A \cap B) \setminus C] \cup [(A \cap C) \setminus B] \\ &= [(A \cap B) \setminus (A \cap C)] \cup [(A \cap C) \setminus (A \cap B)] \\ &= (A \cap B) \triangle (A \cap C) \end{aligned}$$

Finally, let $x \in A \triangle (A \triangle B)$. Then $x \in A \setminus (A \triangle B)$ or $x \in (A \triangle B) \setminus A$. If $x \in A \setminus (A \triangle B)$ then $x \in A$ and $x \notin A \cup B$ but $x \in A \cap B$ In any case $x \in B$. If $x \in (A \triangle B) \setminus A$ then $x \in (A \setminus B) \cup (B \setminus A)$ and $x \notin A$. Thus $x \notin A$ and so $x \in B \setminus A$ so $x \in B$. Hence $A \triangle (A \triangle B) \subseteq B$. On the other hand if $x \in B$ then either $x \in A \cap B$ or $x \in B \setminus (A \cap B)$. In the first case $x \in A$ and $x \notin (A \cup B) \setminus (A \cap B)$ so that $x \in A \setminus (A \triangle B)$. In the second case, $x \in B \setminus (A \cap B)$ so $x \in B \setminus A$ and hence $x \in (A \triangle B) \setminus A$. In either case $x \in A \triangle (A \triangle B)$, which proves the equality.

1.2.1. Solution. Yes.

1.2.3. Solution. Yes.

1.2.5. Solution. f is not injective but is surjective.

1.2.7. Solution. Yes f is injective but not surjective.

1.2.9. Solution. Yes f is injective but not surjective.

1.2.11. Solution. Easy

1.2.13. Solution. This is false.

1.2.15. Solution. This is true. First we show that $f^{-1}(U \cap V) \subseteq f^{-1}(U) \cap f^{-1}(V)$ and to this end let $x \in f^{-1}(U \cap V)$. Then $f(x) \in U \cap V$, so $f(x) \in U$ and $f(x) \in V$. Then $x \in f^{-1}(U)$ and $x \in f^{-1}(V)$ so that this part of the result follows. Conversely, if $x \in f^{-1}(U) \cap f^{-1}(V)$ then $x \in f^{-1}(U)$ and $x \in f^{-1}(V)$.Thus $f(x) \in U$ and $f(x) \in V$, so that $f(x) \in U \cap V$. Hence $x \in f^{-1}(U \cap V)$. Thus $f^{-1}(U) \cap f^{-1}(V) \subseteq f^{-1}(U \cap V)$ and the result follows.

1.2.17. Solution. This is true.

1.2.19. Solution. We do this somewhat informally to give the idea. Let $A_1, A_2, \ldots, A_n \ldots$ be countably many countable sets and let $A = \cup_{i \geq 1} A_i$. We list the elements of A_i and write

$$A_i = \{a_{i,1}, a_{i,2}, a_{i,3}, \ldots, a_{ij}, \ldots\}.$$

If we now list these sets in the form of a matrix, then a listing of the entire set will become evident:

$$\begin{array}{ccccc} a_{11} & a_{12} & \ldots & a_{1n} & \ldots \\ a_{21} & a_{22} & \ldots & a_{2n} & \ldots \\ a_{31} & a_{32} & \ldots & a_{3n} & \ldots \\ \vdots & \vdots & \ddots & \vdots & \end{array}$$

We can now form the listing of A as

$$\{a_{11}, a_{12}, a_{21}, a_{31}, a_{22}, a_{13}, a_{14}, a_{23}, a_{32}, a_{41}, a_{51}, \ldots\}$$

1.3.1. Proof. Suppose that A is infinite. Then A contains a countable subset B. Let $B = \{b_n \mid n \in \mathbb{N}\}$. Define the mapping $\sigma_n : A \longrightarrow A$ by: $\sigma_n(b_n) = b_{n+1}, \sigma_n(b_{n+1}) = b_n$, and $\sigma_n(a) = a$ whenever $a \notin \{b_n, b_{n+1}\}$. It is not hard to prove that σ_n is a permutation of A for every $n \in \mathbb{N}$ (indeed σ_n is a transposition) and $\sigma_n \neq \sigma_k$ whenever $n \neq k$. It follows that $\mathbf{S}(A)$ is infinite.

Conversely, suppose that $\mathbf{S}(A)$ is infinite. If we suppose that A is finite and $|A| = n$, then as above, $|\mathbf{S}(A)| = n!$. This contradiction shows that A must be infinite.

1.3.3. Solution. Yes.

1.3.5. Solution. Yes, it is injective.

So it is easy to conclude that the inverse function will look like:

$$f^{-1}(x) = \begin{cases} \frac{x}{2} \text{ whenever } x \geq 0, \\ \frac{x}{4} \text{ whenever } x < 0. \end{cases}$$

1.3.7. Solution. This mapping is not injective, so there cannot be an inverse.

1.3.9. Solution. We have $(g \circ f)(x) = (x^2/2) - 1$;
$(f \circ g)(x) = (x^2/4) - 2x + 6$;
$((f \circ g) \circ f)(x) = x^4/4 - x^2 + 3$;
$(f \circ (g \circ f))(x) = x^4/4 - x^2 + 3$.

1.3.11. Solution. Suppose that f, g are injective. Let $(g \circ f)(a) = (g \circ f)(b)$, where $a, b \in A$.Then $g(f(a)) = g(f(b))$. Since g is injective it follows that $f(a) = f(b)$ and then since f is injective we have $a = b$, so that $g \circ f$ is injective. Likewise, if f, g are surjective then let $c \in C$. There is an element $b \in B$ such that $g(b) = c$ since g is surjective. However f is surjective so there is an element $a \in A$ such that $f(a) = b$. Then $g(f(a)) = g(b) = c$ which shows that $g \circ f$ is surjective.

1.3.13. Solution.
$(1\,3\,5\,7\,8)(2\,6\,9\,10)(4\,11) = (1\,3)(1\,5)(1\,7)(1\,8)(2\,6)(2\,9)(2\,10)(4\,11)$.

1.3.15. Solution. Even.

1.3.17. Solution. Let $\alpha = (a_1\, a_2\, a_3\, \ldots a_r)$. Then

$$\alpha(a_s) = \begin{cases} a_{s+1} \text{ if } s < r \\ a_1 \text{ if } s = r \end{cases} .$$

Thus if $k < r$ we have $\alpha^k(a_1) = a_{1+k} \neq a_1$ so α^k is not the identity map. On the other hand $\alpha^r(a_s) = a_s$ for all s and since α fixes x if $x \notin \{a_1, \ldots, a_r\}$ it follows that α^r is the identity map. The result follows.

1.3.19. Solution. $(a_1\, a_k\, a_{k-1}\, a_{k-2} \ldots a_3\, a_2)$.

1.4.1. Solution. Let $A = [a_{jk}]_{1 \leq j,k \leq n}, B = [b_{jk}]_{1 \leq j,k \leq n}$. Then $a_{jk} = 0$ whenever $j \neq k$. We have $AB = \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} & \ldots & a_{11}b_{1n} \\ a_{22}b_{21} & a_{22}b_{22} & \ldots & a_{22}b_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{nn}b_{n1} & a_{nn}b_{n2} & \ldots & a_{nn}b_{nn} \end{pmatrix}$, $BA = \begin{pmatrix} a_{11}b_{11} & a_{22}b_{12} & \ldots & a_{nn}b_{1n} \\ a_{11}b_{21} & a_{22}b_{22} & \ldots & a_{nn}b_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{11}b_{n1} & a_{22}b_{n2} & \ldots & a_{nn}b_{nn} \end{pmatrix}$. It follows that $a_{11}b_{12} = a_{22}b_{12}$ or

$(a_{11} - a_{22})b_{12} = 0$, which implies that $b_{12} = 0$. Similarly $b_{jk} = 0$ whenever $j \neq k$.

1.4.3. Solution. If we interchange rows m and t of the matrix A then the m th and t th rows of AB are also interchanged.

1.4.5. Solution.

$$\begin{pmatrix} 1 & 3 & 6 & 10 \\ 0 & 1 & 3 & 6 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

1.4.7. Solution.

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 3 & 3 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

1.4.9. Solution. $\begin{pmatrix} 1 & 4 & 0 & 0 \\ 0 & 1 & 4 & 0 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 1 \end{pmatrix}$.

1.4.11. Solution.
Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then $\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ so that $\begin{pmatrix} a & 0 \\ c & 0 \end{pmatrix} = \begin{pmatrix} a & b \\ 0 & 0 \end{pmatrix}$. This implies that $b = c = 0$. Also $\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}$ from which we deduce that $a = d$. Hence $A = a\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. However any matrix of this form has the property that $AX = XA$ since $IX = XI = X$. In general the result is the same: If $A \in \mathbf{M}_n(\mathbb{R})$ satisfies $AX = XA$ for all matrices in $\mathbf{M}_n(\mathbb{R})$ then $A = rI$ for some real number r.

1.4.13. Solution.

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

1.4.15. Solution. We can see that $(AB)^{-1} = B^{-1}A^{-1}$.

1.4.17. Solution. We get the possibilities:

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

If $b \neq 0$ then $A = \begin{pmatrix} \alpha & b \\ c & -\alpha \end{pmatrix}$, where $\alpha = \pm\sqrt{-1-bc}$, where we assume that $-1-bc \geq 0$ of course.

1.4.19. Solution. First note that $(A+B)^t = A^t + B^t$ and that $(A^t)^t = A$. Indeed, let $A = [a_{ij}], B = [b_{ij}]$ and $C = [c_{ij}]$, where $C = A+B$. If $D = (A+B)^t$ and $D = [d_{ij}]$ then $d_{ij} = c_{ji} = a_{ji} + b_{ji} = e_{ij} + f_{ij}$, where $A^t = [e_{ij}], B^t = [f_{ij}]$ and it therefore follows that $(A+B)^t = A^t + B^t$. Then $(A+A^t)^t = A^t + (A^t)^t = A^t + A = A + A^t$. Also $(A-A^t)^t = A^t - (A^t)^t = A^t - A = -(A-A^t)$. Thus the claims concerning $A+A^t$ and $A-A^t$ follow.

1.5.1. Solution. Not associative. No identity element.

1.5.5. Solution. Associative. Commutative. The pair $(1,0)$ is an identity element.

1.5.7. Answer.

	e	a	b	c
e	e	a	b	c
a	a	b	c	e
b	b	c	e	a
c	c	e	a	b

.

1.5.9. Solution. (a) Not reflexive.
(b) Symmetric.
(c) Transitive and not an equivalence relation.

1.5.11. Solution. Operation is not associative but is commutative.

1.5.13. Solution. We note that $a \equiv b \pmod{m}$ if and only if $b - a = km$ for some integer k. Thus $a \equiv a \pmod{m}$ since $a - a = 0$ is divisible by m. Also if $a \equiv b \pmod{m}$ then there is an integer k such that $b - a = km$ so $a - b = (-k)m$, for the integer $-k$, and hence $b \equiv a \pmod{m}$. Finally if $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$ then there are integers k, l such that $b - a = km, c - b = lm$. Then $c - a = (c-b) + (b-a) = lm + km = (l+k)m$, for the integer $l+k$, so $a \equiv c \pmod{m}$.

The equivalence classes of 0 modulo 7 and 5 are respectively

$$\{0, \pm 7, \pm 14, \pm 21, \dots\} \text{ and } \{0, \pm 5, \pm 10, \dots\}.$$

Clearly these sets are different.

1.5.15. Solution. We deal first with ▼: Not associative, Not commutative. Also ▲ is not associative but is commutative. Finally ■ is also not associative but is commutative. Next note that ▼ does not have an identity associated with it. Likewise ▲ has no identity. Likewise ■ has no identity.

1.5.17. Solution.
(1) Does not hold.
(2) Does not hold.
(3) Does not hold.
(4) Does hold.
(5) Does hold.

1.5.19. Solution.
(a) First $(a,b) \simeq (a,b)$ since $a^2+b^2 = a^2+b^2$. Hence $\simeq$ is reflexive. Next if $(a,b) \simeq (c,d)$ then $a^2+b^2 = c^2+d^2$ so $c^2+d^2 = a^2+b^2$ and hence $(c,d) \simeq (a,b)$. Thus $\simeq$ is symmetric. Finally if $(a,b) \simeq (c,d)$ and $(c,d) \simeq (e,f)$, then we have $a^2+b^2 = c^2+d^2 = e^2+f^2$ so $a^2+b^2 = e^2+f^2$ and $(a,b) \simeq (e,f)$, which means that $\simeq$ is transitive.
(b) $\{(0,0)\}$.
(c) $(x,y) = (1,0),(0,1),(1/2,\sqrt{3}/2),(-1,0),(0,-1)$ are all equivalent to $(1,0)$.

CHAPTER 2

2.1.1. Proof. This result is true for $n = 1$.
Next assume this equation holds for $k < n$ so
$2+2^2+2^3+2^4+\cdots+2^k = 2^{k+1}-2$.
Then by induction hypothesis $2+2^2+2^3+2^4+\cdots+2^n = (2+2^2+2^3+2^4+\ldots 2^{n-1})+2^n = (2^n-2)+2^n = 2\cdot 2^n-2 = 2^{n+1}-2$. The Principle of Mathematical Induction now demostrates the result.

2.1.3. Proof. Use induction on n. The induction starts at $n = 1$. Assume the formula is true for all $k < n$. Then

$$1+4+9+\cdots+(n-1)^2 = \frac{(n-1)n(2(n-1)+1)}{6}.$$

Based on this induction hypothesis we prove our equation.

$$\begin{aligned} 1+4+9+\cdots+(n-1)^2+n^2 &= \frac{(n-1)n(2(n-1)+1)}{6}+n^2 \\ &= \frac{n(n+1)(2n+1)}{6}. \end{aligned}$$

The result follows by the Principle of Mathematical Induction.

2.1.5. Solution. True for $n = 3$. Suppose that $n \geq 3$. The induction hypothesis is $2^k > 2k+1$ for all $3 \leq k < n$. Put $t = n-1$; then $n = t+1$. We have

$2^n - 2n - 1 = 2^{t+1} - 2t - 2 - 1$. We observe that $-2t - 3 > -4t$ if $t \geq 2$, and therefore $2^{t+1} - 2t - 2 - 1 > 2^{t+1} - 4t = 2(2^t - 2t) \geq 0$, when $t \geq 2$.

2.1.7. Proof. The formula holds for $n = 1$.
The induction hypothesis states that the following equation holds.
$1 \cdot 1! + 2 \cdot 2! + 3 \cdot 3! + \cdots + (n-1) \cdot (n-1)! = n! - 1$.
Then we have $1 \cdot 1! + 2 \cdot 2! + 3 \cdot 3! + \cdots + (n-1) \cdot (n-1)! + n \cdot n! = n! - 1 + n \cdot n! = n!(n+1) - 1 = (n+1)! - 1$.

2.1.9. Proof. If $n = 1$ this inequality is obvious. We have the following induction hypothesis,

$$1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \frac{1}{\sqrt{4}} + \cdots + \frac{1}{\sqrt{n-1}} \leq 2\sqrt{n-1}$$

Now

$$1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \frac{1}{\sqrt{4}} + \cdots + \frac{1}{\sqrt{n}}$$
$$= 1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \frac{1}{\sqrt{4}} + \cdots + \frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{n}} \leq 2\sqrt{n-1} + \frac{1}{\sqrt{n}}.$$

However $2\sqrt{n-1} + \frac{1}{\sqrt{n}} \leq 2\sqrt{n}$. The result follows by the Principle of Mathematical Induction.

2.2.1. Solution. It follows since $n^2 + n = n(n+1)$.

2.2.3. Proof. Among the numbers $n, n+1, n+2$ at least one is even and one is divisible by 3.

2.2.5. Proof. The product is $n(n+1)(n+2)(n+3)(n+4)$, where $n \geq 1$.
As we proved in Exercise 2.2.3, this product is divisible by 3. Among these consecutive numbers $n, n+1, n+2, n+3, n+4$ there are at least two even consecutive integers, one of which must be divisible by 4. So their product is divisible by 8.
Note also that among five consecutive integers there is one divisible by 5.

So our product $n(n+1)(n+2)(n+3)(n+4)$ is divisible by the product of the relatively prime divisors 3, 8, and 5, which is equal to 120.

2.2.7. Solution. Write $n = x + 10y$, where x, y are integers and $0 \leq x, y \leq 9$. Furthermore, $n = 4(x+y) + 3$ and $n = 3xy + 5$. When we solve these we obtain $x = 3$ and $n = 23$.

2.2.9. Proof. Write $k^{73}-k^{37}=k^{37}(k-1)(k^2+k+1)(k^6+k^3+1)(k+1)(k^2-k+1)(k^6-k^3+1)(k^{18}+1)$. Show this is divisible by $2,3,5$.

Now, if k divisible by 5, then k^{37} is divisible by 5, so the product above is divisible by 5. If the remainder upon dividing k by 5 is 1, then $k-1$ is divisible by 5, and the product above is divisible by 5. Continue in this fashion.

2.2.11. Solution. (i) We have $b=ad$ and $c=bu$ for some $d,u\in\mathbb{Z}$. Then $ca(du)$, so that $a \mid c$, since $du\in\mathbb{Z}$.
(iii) We have again $b=ad$. Then $bc=(ac)d$, so that $ac \mid bc$.
(v) We have $b=au$ and $d=cv$ for some $u,v\in\mathbb{Z}$. Then

$$bd=(ac)(uv),$$

so that $ac \mid bd$.

2.2.13. Solution. Suppose that $c=qd+r$, where q,r are integers and $0\le r<d$. Then $r=c-qd$. By definition there are integers s,t such that $a=dt$, $b=ds$. Since $c=au+bv$ we have $r=au+bv-qd=dtu+dsv-qd=d(tu+sv-q)$. Since $tu+sv-q$ is an integer we have that d divides r and hence $r=0$. Therefore d divides c.

2.2.15. Solution. Let d be the greatest common divisor of a and b and write $a=du, b=dv$ for integers u,v. Then a divides ab/d. Likewise b divides ab/d. Next suppose that a and b both divide x, say $x=sb=at$ for integers s,t. Then $a|sb$ so du divides dvs and hence u divides vs. However u and v are relatively prime so u divides s. Then $x=sb=sdv=urdv$ for some integer r. This means that x is divisible by $udv=ab/d$. Hence the result.

2.2.17. Solution. Let $d_1=\mathbf{GCD}(a,b), d_2=\mathbf{GCD}(d_1,c), e_1=\mathbf{GCD}(b,c), e_2=\mathbf{GCD}(a,e_1)$. We need to prove that $d_2=e_2$. To this end we prove first that d_2 divides e_2. Thus we first prove that d_2 divides e_1 and that d_2 divides a. Since d_2 divides d_1 we have $d_1=rd_2$ for some $r\in\mathbb{Z}$ and since d_1 divides a we have $a=sd_1$ for some $s\in\mathbb{Z}$. Hence $a=srd_2$, so d_2 divides a. Also d_2 divides c, by definition. Since d_1 divides b we have $b=td_1$ for some $t\in\mathbb{Z}$. Then $b=trd_2$ so d_2 divides b. Then because d_2 divides b and c it divides their greatest common divisor which is e_1. Hence d_2 divides both e_1 and a so it divides their greatest common divisor which is e_2. By symmetry we have e_2 divides d_2 and this implies that $e_2=d_2$.

2.2.19. Solution. There are integers m,n,r,s such that $ma+nc=1, rb+sc=1$. Hence $(ma+nc)(rb+sc)=1$. Thus $(mr)(ab)+(nrb+mas+ncs)c=1$.

2.3.1. Solution. We know $x^2+2x-3=(x-1)(x+3)$. We need to find all natural numbers x such that one of these factors is 1. There is only one natural number that satisfies our conditions, namely $x=2$.

2.3.3. Solution. We have $x^2-5x+6=(x-2)(x-3)$. There is only one natural number that satisfies our conditions, namely $x=4$.

2.3.5. Solution. $a=2^4\cdot 3^2\cdot 5^2\cdot 7\cdot 11\cdot 13\cdot 17\cdot 19\cdot 23$.

2.3.7. Solution. The obvious example of such a triple is $3,5,7$. There is no other triple like this.

2.3.9. Solution. Let n be an odd integer greater than 6. Take a prime $2\neq p<n$. Then $n-p$ is an even number. Since $n>6$, we can choose p such that $n-p>2$. By the Goldbach conjecture, $n-p=a+b$, where a,b are prime and then $n=a+b+p$.

2.3.11. Solution. When $0<i<p$, both $i!$ and $(p-i)!$ are relatively prime to p, since p is prime. Hence $i!(p-i)!$ is also relatively prime to p. However $p!$ is divisible by p. Hence $\binom{p}{i}$ is divisible by p.

2.3.13. Solution. For example $5,13,17$. Next suppose that k,l are natural numbers and consider the product $(4k+1)(4l+1)=16kl+4l+4k+1=4(4kl+l+k)+1$ which is of the form $4z+1$, where z is a natural number. Next we are supposed to prove that a product of n integers, where $n\geq 1$, of the form $4k+1$ is again of this same form. The induction step is similar.

2.3.15. Solution. Note that $6k+2,6k+3,6k+4,6k$ are all composite, if $k\geq 1$. Thus every prime number at least 5 is of the form $6k+1$ or $6k+5$. Note that $(6k+1)(6l+1)=6(6lk+l+k)+1$ so a product of two numbers of the form $6r+1$ again has this form.

2.3.17. Solution. 116 and 640.

2.4.1. Proof. Let a be an odd number. Then $b=\frac{a^2-1}{2}$, and $c=\frac{a^2-1}{2}+1$. Then $a^2+b^2=c^2$.

2.4.3. Proof. We need to prove that this decimal is nonrepeating. Assume the contrary. Let $n,n+1,\ldots,n+m$ be consecutive natural numbers that correspond to the repeated segment in the decimal representation of our number, Thus

$$x=0.12345678\ldots(n-1)\,\overline{n\,(n+1)\ldots(n+m)}.$$

But according to the pattern

$$x = 0.12345678\ldots(n-1)n(n+1)\ldots(n+m)(n+m+1)\ldots.$$

Since $n \neq n+m+1$, we obtain a contradiction.

2.4.5. Proof. Prove that there is no rational number $r = \frac{p}{q}$, such that $(\frac{p}{q})^2 = 5$. Assume that there exists a rational number r such that $r^2 = 5$. Assume that p and q in the representation $r = \frac{p}{q}$ are relatively prime, so have no common divisors. Then $(\frac{p}{q})^2 = 5$.

It follows that $5q^2 = p^2$ so 5 divides p. Likewise q is also divisible by 5. But then the fraction $\frac{p}{q}$ is reducible, so is not in its lowest terms. So $\sqrt{5}$ is irrational.

2.4.7. Proof. Similar to the proof of 2.4.5.

2.4.9. Proof. Similar to the proof of 2.4.5.

2.4.11. Solution. Suppose that (a,b) is related to (c,d) and (c,d) is related to (e,f), where a,b,c,d,e,f are integers and $b,d,f \neq 0$. Then $ad = bc$ and $cf = de$. Hence $af = be$. This shows that (a,b) is related to (e,f).

2.4.13. Solution.
If $r = 0$ there is nothing to prove so assume that $r \neq 0$. There is a polynomial $a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$, where the coefficients a_i are rational, having α as a root. Thus

$$a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_n\alpha^n = 0.$$

Then

$$a_0 + \frac{a_1}{r}(r\alpha) + \frac{a_2}{r^2}(r\alpha)^2 + \cdots + \frac{a_n}{r^n}(r\alpha)^n = 0$$

and each of the terms $\frac{a_i}{r^i}$ is rational. Clearly $r\alpha$ is a root of the polynomial

$$a_0 + \frac{a_1}{r}X + \frac{a_2}{r^2}X^2 + \cdots + \frac{a_n}{r^n}X^n$$

with rational coefficients.

CHAPTER 3

3.1.1. Solution. Note that $\left(a+bi\sqrt{5}\right)\left(a_1+b_1i\sqrt{5}\right) = (aa_1-5bb_1) + (ab+ba_1)\,i\sqrt{5} \in G$. Thus the operation is a binary operation on G.

Multiplication is an associative operation in the set of complex numbers, as is well-known. The identity is $1 = 1+0i\sqrt{5} \in G$.

The inverse of $z = a+bi\sqrt{5}$ should be $w = \frac{1}{a+bi\sqrt{5}} = \frac{a-bi\sqrt{5}}{a^2+5b^2}$.

3.1.3. Solution.

$\times$	e	a	b	c
e	e	a	b	c
a	a	b	c	e
b	b	c	e	a
c	c	e	a	b

.

3.1.5. Solution. First of all check whether this set is closed under the operation of multiplication.

Multiplication is an associative operation in the set of complex numbers

The identity $1 = 1+0i$ is an element of the set G.

The inverse element of $\alpha = a+bi$ is $\alpha^{-1} = a-bi$ Yes, this set is a group.

3.1.7. Proof. Indeed, if $a,b \in G$, then $(ab)^2 = abab = 1$, and $ab = (ab)^{-1} = b^{-1}a^{-1}$. Since $g^2 = 1$ for each element g of G, $g = g^{-1}$. It follows from the last equation that $ab = ba$.

3.1.9. Solution. The set G is obviously closed under this operation.

The associative property holds since $[(m,a)\star(n,b)]\star(x,y) = (m+an,ab)\star(x,y) = (m+an+abx,aby)$, and $(m,a)\star[(n,b)\star(x,y)] = (m,a)\star(n+bx,by) = (m+an+abx,aby)$.

The identity is $(0,1)$.

Thus the inverse of (n,b) is $(-nb,b)$. Hence G is a group with this operation $\star$.

The operation is not commutative.

3.1.11. Solution. $1,-1$ are the only generators of $\mathbb{Z}$.

3.1.13. Solution.
(a) $x,x^5,x^7,x^{11},x^{13},x^{17},x^{19},x^{23}$.
(b) x,x^2,x^3,x^4,x^5,x^6.
(c) $x,x^5,x^7,x^{11},x^{13},x^{17}$.

3.1.15. Solution. First we note that certainly $a\circ b$ is a real number in A.
Next the associative property holds: Let $a,b,c \in A$. Then $(a\circ b)\circ c = a+b+c+ab+ac+bc+abc$. Likewise $a\circ(b\circ c) = a+b+c+ab+ac+bc+abc$ so the two expressions are equal. This proves associativity.

The real number $0 \in A$ is the identity. Also when $a \neq -1$ then $\frac{-a}{1+a} \in A$ exists The inverse of a is $\frac{-a}{1+a}$. Hence A is a group under the operation $\circ$. Furthermore it is abelian since $a \circ b = b \circ a$, for all $a, b \in A$.

3.1.17. Solution.
The Cayley table for $X(7)$ is

$\cdot$	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	1	3	5
3	3	6	2	5	1	4
4	4	1	5	2	6	3
5	5	3	1	6	4	2
6	6	5	4	3	2	1

The Cayley table for $X(8)$ is

$\cdot$	1	3	5	7
1	1	3	5	7
3	3	1	7	5
5	5	7	1	3
7	7	5	3	1

Here $X(7)$ is cyclic since the element 3, for example, has order 6. On the other hand $X(8)$ is not cyclic since the non-identity elements each have order 2.

3.1.19. Solution. 8.

3.2.1. Solution. The elements x are $g, g^3, g^5, g^7, g^9, g^{11}, g^{13}, g^{15}$.

3.2.3. Solution. Each of these matrices has determinant ± 1 and it is easy to see that the inverse of each of the matrices in N is also in N. It is also easy to see that the product of any two of the matrices in N is also in N. It is therefore easy to check directly that these matrices form a subgroup of $\mathbf{GL}_2(\mathbb{Z})$. N is not a normal subgroup.

3.2.5. Solution. . It is clear that H is nonempty and that if $\alpha, \beta \in H$ then $\alpha\beta^{-1} \in H$. It is easy to see that there is a bijective mapping φ from H to $\mathbf{A}_4$ defined by the rule:

$\varphi(\pi) = \phi$, where ϕ is the element of $\mathbf{A}_4$ which is obtained from π by eliminating the last column $\begin{pmatrix} 5 \\ 5 \end{pmatrix}$. Since the map is bijective, H is a subgroup having the same order as A_4, so $|H| = 12$.

3.2.7. Solution. The order of this element is 2.

3.2.9. Proof. Indeed, this group can only have subgroups whose orders are divisors of p^2, namely $1, p$, of p^2.

Since the order of the identity subgroup is 1, and the order of the group itself is p^2, we just need to prove that there is only one subgroup of this group having order p. Suppose now that $G = \langle g \rangle$. Clearly $\langle g^p \rangle$ has order p. If g^r is some other element of order p then $g^{rp} = e$. By Lagrange's Theorem, p^2 divides rp and therefore p divides r. Hence $g^r \in \langle g^p \rangle$ so we must have $\langle g^r \rangle = \langle g^p \rangle$.

3.2.11. Solution. Suppose that $\alpha \in Z(\mathbf{S}_n)$ and that $\alpha = \alpha_1 \alpha_2 \dots \alpha_k$, as a product of disjoint cycles. Suppose one of these cycles, α_1 say, is of length at least 3 and let $\alpha_1 = (u_1 u_2 u_3 \dots u_k)$. Consider $\alpha(u_1 u_2) = (u_1 u_2)\alpha$, since $\alpha \in Z(\mathbf{S}_n)$. Since $(u_1 u_2)$ must commute with $\alpha_2, \dots, \alpha_k$ it follows that $\alpha_1(u_1 u_2) = (u_1 u_2)\alpha_1$. However when we compute both sides of this product we obtain, respectively $(u_2 u_3 \dots u_k)$ and $(u_1 u_3 \dots u_k)$. Since these are not equal we obtain a contradiction. Thus each of the cycles $\alpha_1, \dots, \alpha_k$ is a transposition. So $\alpha_1 = (u_1 u_2)$. If $k = 1$, so that $\alpha = \alpha_1$, then pick $u_3 \neq u_1, u_2$ and observe that $(u_1 u_2)(u_1 u_3) = (u_1 u_2 u_3)$ whereas $(u_1 u_3)(u_1 u_2) = (u_1 u_3 u_2)$. These are not equal. Thus $k > 1$. Now let $\alpha_1 = (u_1 u_2), \alpha_2 = (u_3 u_4)$. Then since $\alpha \in Z(\mathbf{S}_n)$ we must have

$$(u_1 u_2)(u_3 u_4)(u_1 u_3) = (u_1 u_3)(u_1 u_2)(u_3 u_4).$$

However the first of these is $(u_1 u_2 u_3 u_4)$ and the second is $(u_1 u_4 u_3 u_2)$, which are clearly not equal. This proves the result.

3.2.13. Solution. Note that $g^{-1} h^n g = (g^{-1} h g)^n$.

3.2.15. Solution. We first show that X is a subgroup of $\mathbf{SL}_2(\mathbb{Z})$. Of course the identity matrix is an element of X so X is non-empty. The easiest way to proceed now is to note that if $A = \begin{pmatrix} 1+na & nb \\ nc & 1+nd \end{pmatrix}$ has determinant 1 then its inverse is $\begin{pmatrix} 1+nd & -nb \\ -nc & 1+na \end{pmatrix}$, which is clearly a matrix of the same type as A, so $A^{-1} \in X$. Next we note that in fact any element of X can be written as $A = I + nU$ where U is an integer matrix. If $B = I + nV$ for some integer matrix V then

$$AB = (I+nU)(I+nV) = I + nU + nV + n^2 UV = I + n(U + V + nUV)$$

and this again is clearly an element of X. Hence X is a subgroup. Next if C is any arbitrary matrix in $\mathbf{SL}_2(\mathbb{Z})$ then we have

$$C^{-1}AC = C^{-1}(I+nU)C = C^{-1}C + nC^{-1}UC = I + nC^{-1}UC.$$

Of course $\mathbf{det}\, C^{-1}AC = \mathbf{det}\, C^{-1} \cdot \mathbf{det}\, A \cdot \mathbf{det}\, C = 1$, so that $C^{-1}AC \in X$.

3.2.17. Solution. Since the index of $\langle\alpha\rangle$ in $\mathbf{D}_{2n}$ is 2, it is clear that $\langle\alpha\rangle$ is a normal subgroup of $\mathbf{D}_{2n}$. To show that $\langle\beta\rangle$ is not normal it is best to compute the cosets $\alpha\langle\beta\rangle$and $\langle\beta\rangle\alpha$. It is easy to see that these cosets are different because $\alpha\beta \neq \beta\alpha$.

3.2.19. Solution. Let $g \in G$. Each of the sets $g^{-1}Hg$ is a subgroup of G: Clearly $e = g^{-1}eg \in g^{-1}Hg$. Also if $h,k \in g^{-1}Hg$ then $h = g^{-1}xg$ and $k = g^{-1}yg$ for some $x,y \in H$. Then $hk^{-1} = (g^{-1}xg)(g^{-1}y^{-1}g) = g^{-1}xy^{-1}g \in g^{-1}Hg$ since H is a subgroup of G. It follows that N is also a subgroup of G since the intersection of an arbitrary collection of subgroups is again a subgroup. To show that N is normal we have to show that if $x \in G$ and $n \in N$ then $x^{-1}nx \in N$. However if $g \in G$ then $n \in xg^{-1}H(xg^{-1})^{-1}$ so $n = xg^{-1}hgx^{-1}$ for some $h \in H$. Hence $x^{-1}nx = x^{-1}xg^{-1}hgx^{-1}x = g^{-1}hg \in g^{-1}Hg$ and this is true for all $g \in G$. Hence $x^{-1}nx \in N$, as needed.

3.3.1. Proof. Easy

3.3.3. Solution. Let $x,y \in \mathbb{N}$. Consider $\Theta(xy)$. If the number xy is odd, then both numbers x and y are odd, and $\Theta(xy) = 1 = \Theta(x)\Theta(y)$. If one of the numbers is even, then the product is even, and $\Theta(xy) = 0 = \Theta(x)\Theta(y)$.

3.3.5. Proof. This follows because an element of order r is mapped by a homomorphism to an element of order s where s divides r.

3.3.7. Proof. Let $x = hn$ and $y = h_1n_1$ be elements of HN where $h,h_1 \in H, n,n_1 \in N$. Then $y^{-1} = (h_1n_1)^{-1}$. Consider $xy^{-1} = hn(h_1n_1)^{-1} = hnn_1^{-1}h_1^{-1} = h\left(h_1^{-1}h_1\right)nn_1^{-1}h_1^{-1} = hh_1^{-1}(h_1nn_1^{-1}h_1^{-1})$. Let $m = nn_1^{-1} \in N$. Since $N \lhd G$, the conjugate $h_1nn_1^{-1}h_1^{-1} = h_1mh_1^{-1} = n_2 \in N$. So $hh_1^{-1}h_1nn_1^{-1}h_1^{-1} = hh_1^{-1}n_2 = h_2n_2 \in HN$, where $h_2 = hh_1^{-1}$. Thus HN is a subgroup of G.

3.3.9. Proof. Assume the contrary. Let L/N be a proper normal subgroup of G/N. Then $N < L \lhd G$ and $G \neq L$, a contradiction.

Of course, the converse statement is also correct. Indeed, suppose that there is no proper normal subgroup in G/N. If we assume that N is not a maximal normal subgroup in G, we can choose a normal

proper subgroup L of G such that $N < L \lhd G$. But then $L/N \lhd G/N$ will be a proper normal subgroup in G/N, a contradiction.

3.3.11. Solution. First we note that HN is a subgroup of G, using Problem 3.3.7. Also N is a normal subgroup of HN. We next define a function $\theta : H \longrightarrow HN/N$ by $\theta(h) = hN$ for all $h \in H$. Then use the First Isomorphism Theorem.

3.3.13. Solution. First we note that M/N is a normal subgroup of G/N, which can be easily verified. We define a function $\theta : G/N \longrightarrow G/M$ by $\theta(gN) = gM$, for all $g \in G$. The First Isomorphism Theorem now implies the result.

3.3.15. Solution. Use Lagrange's Theorem.

3.3.17. Solution. Let $x \in N, y \in M$. Then $x^{-1}yx \in M$ so $y^{-1}x^{-1}yx \in M$ also. Similarly $y^{-1}x^{-1}y \in N$ so $y^{-1}x^{-1}yx \in N$. Hence $y^{-1}x^{-1}yx \in M \cap N = \{e\}$. Thus $y^{-1}x^{-1}yx = e$ and it follows that $yx = xy$, as required.
For the example note that $\mathbf{S}_3$ has the two subgroups $\langle (1\,2) \rangle$ and $\langle (1\,2\,3) \rangle$.

3.3.19. Solution. We have to think of $\mathbb{Z}[X]$ additively and $\mathbb{Q}^+$ multiplicatively. We have

$$\begin{aligned} &f((a_0 + a_1X + \cdots + a_nX^n) + (b_0 + b_1X + \cdots + b_nX^n)) \\ &\quad = f(a_0 + a_1X + \cdots + a_nX^n) f(b_0 + b_1X + \cdots + b_nX^n) \end{aligned}$$

Hence f is a homomorphism.

Use the Fundamental Theorem of Arithmetic to show f is an isomorphism.

CHAPTER 4

4.1.1. Solution. It is obvious that the set R is an abelian group under the given operation of addition. The additive identity is $(0,0)$ and the negative of (a,b) is $(-a,-b)$. The distributive laws hold. The multiplication is associative, again since every product is $(0,0)$. Hence this set is a ring. It is easy to see that every nonzero element of R is a zero divisor, since every product is $(0,0)$. There is no identity element in R again since every product is $(0,0)$.

4.1.3. Solution. The set R is not an abelian group under the given operation of addition.

The distributive laws do not hold either.
So R is not a ring.

4.1.5. Proof. For the element a there is a unique additive inverse $-a$ in R. Adding this element to both sides of our equation and using the associative and commutative properties of addition we have $-a + a + x = -a + b \Longrightarrow (-a+a) + x = -a + b \Longrightarrow 0 + x = -a + b \Longrightarrow x = b - a$. This shows not only that there is a solution but that it is unique.

4.1.7. Solution. This subring consists of the set of elements R of the form $a + b\sqrt{-5}$ where $a, b \in \mathbb{Z}$. The set of all such elements is a ring. Indeed it is a subring of $\mathbb{C}$ since $a+b\sqrt{-5} - (x+y\sqrt{-5}) = (a-x) + (b-y)\sqrt{-5}$ which is again an element of R whenever $a, b, x, y \in \mathbb{Z}$. Likewise we can see that R is closed under multiplication.

There is only one invertible element in this subring: $1 + 0\sqrt{-5}$.

4.1.9. Proof. Let $\begin{pmatrix} a & b \\ -b & a-b \end{pmatrix}, \begin{pmatrix} x & y \\ -y & x-y \end{pmatrix}$ be two elements of L. Then

$$\begin{pmatrix} a & b \\ -b & a-b \end{pmatrix} - \begin{pmatrix} x & y \\ -y & x-y \end{pmatrix} = \begin{pmatrix} a-x & b-y \\ -b+y & a-b-x+y \end{pmatrix}$$

$$= \begin{pmatrix} a-x & b-y \\ -(b-y) & (a-x)-(b-y) \end{pmatrix}$$

so L is a subgroup under addition. Also the product of such matrices is a matrix of the same type. Consider the product of two matrices $\begin{pmatrix} a & b \\ -b & a-b \end{pmatrix}$ and $\begin{pmatrix} a_1 & b_1 \\ -b_1 & a_1-b_1 \end{pmatrix}$. Then

$$\begin{pmatrix} a & b \\ -b & a-b \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ -b_1 & a_1-b_1 \end{pmatrix}$$

$$= \begin{pmatrix} aa_1 - bb_1 & -bb_1 + a_1b + b_1a \\ -a_1b + bb_1 - b_1a & aa_1 - bb_1 - a_1b + bb_1 - b_1a \end{pmatrix}.$$

Since $-(-bb_1 + a_1b + b_1a) = bb_1 - a_1b - b_1a$ and $(aa_1 - bb_1) - (-bb_1 + a_1b + b_1a) = aa_1 - bb_1 + bb_1 - a_1b - b_1a$ this is a ring.
Commutativity follows from the simple observation that all matrix entries in the product are unchanged when we interchange a with a_1 and b with b_1.

4.1.11. Solution. Let $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be an element of the center of the ring $\mathbf{M}_2(\mathbb{R})$. Then in particular we have

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & a \\ 0 & c \end{pmatrix} \text{ and}$$
$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} c & d \\ 0 & 0 \end{pmatrix}.$$

Thus $a = d$ and $c = 0$. Also

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} b & 0 \\ d & 0 \end{pmatrix} \text{ and}$$
$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ a & b \end{pmatrix}.$$

Hence $b = 0$ and $\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$ is our candidate to be an element of the center. However $\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} = aI$ and clearly aI is central. Thus $Z(\mathbf{M}_2(\mathbb{R})) = \{aI | a \in \mathbb{R}\}$.

4.1.13. Solution. Let $R = \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}$. Let $I = \{(a,a,0) | a \in \mathbb{Z}\}$. Let $J = \{(b,0,b) | b \in \mathbb{Z}\}$. Hence $I + J$ is not a subring.

4.1.15. Solution. Note that $(a+b)^2 = a+b$ and $(a+a)^2 = a+a$.

4.1.17. Solution. We can use the ring $\mathbf{M}_2(\mathbb{R})$, for example.

4.1.19. Solution. Consider the function $f : R \longrightarrow R$ defined by $f(x) = ax$.

4.2.1. Solution. These polynomial are equal iff the coefficients of the corresponding powers of X are the same. Then $a = -5, b = -1, c = 6$.

4.2.3. Solution. In this case $g(X) = \pm(X^2 + 6X + 1)$.

4.2.5. Proof. Let $f(X) = a_1X^n + a_2X^{n-1} + \cdots + a_{n-1}X + a_n \in \mathbb{Z}[X]$. Then $f(7) = a_1 7^n + a_2 7^{n-1} + \cdots + a_{n-1}7 + a_n = 11$ and $f(11) = a_1 11^n + a_2 11^{n-1} + \cdots + a_{n-1}11 + a_n = 13$. Subtract the first equation from the second and note that the left-hand side of the equation, $a_1(11^n - 7^n) + a_2(11^{n-1} - 7^{n-1}) + \cdots + a_{n-1}(11 - 7) = 2$, is divisible by 4.

4.2.7. Solution. $a = -1$ or $a = \frac{5}{3}$.

4.2.9. **Solution.** The quotient is $2X^2 - 5X - 11\frac{1}{2}$, the remainder is $-64\frac{1}{2}X + 7\frac{1}{2}$.

4.2.11. **Solution.** We are assuming that $r = a/b$ is a root of $a_0 + a_1X + a_2X^2 + \cdots + a_nX^n = 0$, where $a_i \in \mathbb{Z}$ and $\mathbf{GCD}(a,b) = 1$, so if we substitute in we obtain

$$a_0 + a_1(a/b) + a_2(a/b)^2 + \cdots + a_n(a/b)^n = 0.$$

Obtain

$$a_0b^n + a_1ab^{n-1} + a_2a^2b^{n-2} + \cdots + a_na^n = 0.$$

Then a divides a_0b^n so a divides a_0. Likewise b divides a_na^n so b divides a_n.

4.2.13. **Solution.** The greatest common divisor is $X^2 - X + 1$.
Take $u(X) = 1, v(X) = \frac{6}{7}X - \frac{3}{7}$.

4.2.15. **Solution.** $-1, -2, -3, 4$.

4.2.17. **Solution.** $X^5 + 4X^4 + 7X^3 + 8X^2 + 5X + 2 = (X^2 + X + 1)^2(X + 2)$.

4.2.19. **Solution.** If we think first of this polynomial as a polynomial over $\mathbb{Q}$ then we note that a polynomial over a field of degree 3 either has three linear factors or a linear and an irreducible quadratic factor or is irreducible. By the rational root test applied to this polynomial we see that there are no rational roots. Indeed the only possible rational roots are ± 1 and neither of these is actually a root. It now follows that this polynomial is indeed irreducible over $\mathbb{Q}$ and hence is irreducible over $\mathbb{Z}$.

4.3.1. **Proof.** Since $\mathbb{Q}$ and $\mathbb{Z}$ are rings, it is easy to see that R is a ring. The identity element here is the pair $(1,1)$.

Since $\mathbb{Q}$ is a field, its only ideals are $\mathbb{Q}$ and $\{0\}$. The ideals of the ring $\mathbb{Z}$ are $n\mathbb{Z}$, where n is a natural number or 0. Now it is easy to see that the ideals of R can be written as $\mathbb{Q} \times n\mathbb{Z}$ or $\{0\} \times n\mathbb{Z}$, where $n \in \mathbb{N}_0$.

4.3.3. **Solution.** Certainly $nR \neq \emptyset$. Consider the difference of two elements from nR, say $nx, ny \in nR$, where $x, y \in R$. Then

$$nx - ny = n(x - y) \in nR.$$

Also $nx \cdot y = n(xy) \in R$.
So nR is an ideal.

4.3.5. Solution.

×	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

.

4.3.7. Proof. Certainly $I \neq \emptyset$. Also the difference of each pair of elements of I belongs to I, since to obtain an entry in the difference of two matrices we simply subract the corresponding entries in the matrices concerned. Consider the product $\begin{pmatrix} a & b \\ 0 & c \end{pmatrix}\begin{pmatrix} 0 & d \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & ad \\ 0 & 0 \end{pmatrix} \in I$. Also $\begin{pmatrix} 0 & d \\ 0 & 0 \end{pmatrix}\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} = \begin{pmatrix} 0 & cd \\ 0 & 0 \end{pmatrix}$.

So I is an ideal of R.

4.3.9. Solution. As we mentioned above, the quotient ring $\mathbb{Z}[i]/M$ consists of four cosets $0+M, 1+M, i+M, (1+i)+M$. Since $(1+i)^2+M = 2i+M = 0+M$ it is clear that $1+i+M$ is a zero divisor. There are no other zero divisors.

4.3.11. Solution. Let S be the set of matrices of the form $\begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix}$. Clearly this set is nonempty. Suppose that $\begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix}, \begin{pmatrix} 0 & x \\ 0 & y \end{pmatrix} \in S$ and $\begin{pmatrix} r & s \\ t & u \end{pmatrix} \in \mathbf{M}_2(\mathbb{R})$. We have

$$\begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix} - \begin{pmatrix} 0 & x \\ 0 & y \end{pmatrix} = \begin{pmatrix} 0 & a-x \\ 0 & b-y \end{pmatrix} \in S \text{ and}$$
$$\begin{pmatrix} r & s \\ t & u \end{pmatrix}\begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix} = \begin{pmatrix} 0 & ra+sb \\ 0 & ta+ub \end{pmatrix} \in S.$$

This shows that S is a left ideal. On the other hand

$$\begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix}\begin{pmatrix} r & s \\ t & u \end{pmatrix} = \begin{pmatrix} at & au \\ bt & bu \end{pmatrix}$$

which will not generally be in S. Hence S is not a right ideal.

4.3.13. Solution. We first have to prove that if I is an ideal of R and J is an ideal of S then $I \oplus J$ is an ideal of $R \oplus S$. Clearly $0 \in I, 0 \in J$ so $(0,0) \in I \oplus J$ which is therefore a nonempty subset of $R \oplus S$. Next we observe that if $a, b \in I \oplus J$ then $a = (u,v), b = (x,y)$ for certain elements $u, x \in I, v, y \in J$. Then $a - b = (u,v) - (x,y) = (u-x, v-y)$. This is

an element of $I \oplus J$ since $u - x \in I$ and $v - y \in J$. Next let $(r,s) \in R \oplus S$. Then $(u,v)(r,s) = (ur, vs)$. Since I is an ideal we have $ur \in I$, because $u \in I$ and likewise $vs \in J$. Hence $(u,v)(r,s) \in I \oplus J$. Since also $(r,s)(u,v)$ is also in $I \oplus J$ we have that $I \oplus J$ is an ideal of $R \oplus S$.
If K is an ideal of $R \oplus S$ let

$$I = \{a \in R | (a,0) \in K\}.$$

Show I is an ideal of R. Let

$$J = \{b \in S | (0,b) \in K\}$$

and observe J is an ideal of S. Then show $I \oplus J = K$ as needed.

4.3.15. Solution. The natural map is

$$f\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = \begin{pmatrix} a+p\mathbb{Z} & b+p\mathbb{Z} \\ c+p\mathbb{Z} & d+p\mathbb{Z} \end{pmatrix}.$$

By definition of $\mathbf{GL}_2(\mathbb{Z})$ we must have $ad - bc = \pm 1$ so $(a+p\mathbb{Z})(d+p\mathbb{Z}) - (b+p\mathbb{Z})(c+p\mathbb{Z}) = (ad-bc)+p\mathbb{Z} = \pm 1 + p\mathbb{Z}$ so that the image of $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is indeed an element of $\mathbf{GL}_2(\mathbb{Z}/p\mathbb{Z})$. Next we let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, B = \begin{pmatrix} x & y \\ w & z \end{pmatrix}$ Then

$$AB = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x & y \\ w & z \end{pmatrix} = \begin{pmatrix} ax+bw & ay+bz \\ cx+dw & cy+dz \end{pmatrix}$$

so

$$\begin{aligned} f(AB) &= \begin{pmatrix} ax+bw+p\mathbb{Z} & ay+bz+p\mathbb{Z} \\ cx+dw+p\mathbb{Z} & cy+dz+\mathbb{Z} \end{pmatrix} \\ &= \begin{pmatrix} a+p\mathbb{Z} & b+p\mathbb{Z} \\ c+p\mathbb{Z} & d+p\mathbb{Z} \end{pmatrix}\begin{pmatrix} x+p\mathbb{Z} & y+p\mathbb{Z} \\ w+p\mathbb{Z} & z+p\mathbb{Z} \end{pmatrix} \\ &= f(A)f(B). \end{aligned}$$

Hence f is a homomorphism. The identity of $\mathbf{GL}_2(\mathbb{Z}/p\mathbb{Z})$ is the matrix $\begin{pmatrix} 1+p\mathbb{Z} & 0+p\mathbb{Z} \\ 0+p\mathbb{Z} & 1+p\mathbb{Z} \end{pmatrix}$ so $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a matrix in the kernel if and only if $ad - bc = \pm 1$ and $a \equiv d \equiv 1 \pmod{p}, c \equiv d \equiv 0 \pmod{p}$.

4.3.17. Solution. Let R be a ring with unity in which every subring is an ideal. Show $\{n \cdot 1 | n \in \mathbb{Z}\}$ is a subring of R which must be R.

4.3.19. Solution. S is a nonempty subset of R. Suppose next that $a,b \in S$. Then there are natural numbers n,m such that $a^n = b^m = 0$. Then

$$(a-b)^{n+m} = \sum_{i=0}^{m+n} \binom{m+n}{i} a^i (-b)^{n+m-i},$$

which is 0. Hence $a-b \in S$. Finally if $r \in R$ then $ar \in S$ and since $ar = ra$ it follows that S is an ideal of R.

Next suppose that $a+S$ is an element of R/S that is nilpotent. Then there exists a natural number n such that $(a+S)^n = 0+S$. This means that $a \in S$. Thus $a+S = 0+S$ and this proves the result.

4.4.1. Solution. Prove that the set $\mathbb{R} \times \mathbb{R}$ with the defined operations of addition and multiplication is isomorphic to the field of complex numbers.

For the proof we define a mapping $\alpha : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{C}$ by $\alpha((a,b)) = a+bi$.

It is quite easy to see that this mapping is bijective. Furthermore, if $(a,b),(c,d) \in \mathbb{R} \times \mathbb{R}$ then $\alpha((a,b)+(c,d)) = \alpha((a,b)) + \alpha((c,d))$.

Now definitely

$\alpha((a,b)(a_1,b_1)) = (a+bi)(a_1+b_1i)$. Hence our mapping is an isomorphism. Thus $\mathbb{R} \times \mathbb{R}$ is a ring.

4.4.3. Solution. $K = \mathbb{Z}[\sqrt{3}] = \{a+b\sqrt{3} \mid a,b \in \mathbb{Z}\}$ is a subring of $\mathbb{R}$ because the difference and the product of the numbers from K are clearly elements of K. The difference of matrices from L is also easily seen to be a matrix from L. The product of such matrices also belongs to L. Hence K,L are subrings of their appropriate rings.

We will define the mapping $\alpha : K \longrightarrow L$ by

$$\alpha(a+b\sqrt{3}) = \begin{pmatrix} a & b \\ 3b & a \end{pmatrix}.$$

This mapping is a bijection. To see that this mapping is a homomorphism we have to see that it respects the addition and multiplication. However

$$\begin{aligned} \alpha((a+b\sqrt{3})+(x+y\sqrt{3})) &= \alpha((a+x)+(b+y)\sqrt{3}) \\ &= \begin{pmatrix} a+x & b+y \\ 3(b+y) & a+x \end{pmatrix} \\ &= \begin{pmatrix} a & b \\ 3b & a \end{pmatrix} + \begin{pmatrix} x & y \\ 3y & x \end{pmatrix}, \end{aligned}$$

so the addition is respected. In order to prove that this mapping respects the operation of multiplication we just consider the product $\left(a+b\sqrt{3}\right)\left(x+y\sqrt{3}\right) = (ax+3by)+(ay+bx)\sqrt{3}$. As we can see, the image of this product is $\begin{pmatrix} ax+3by & ay+bx \\ 3(ay+bx) & ax+3by \end{pmatrix}$, which is the product of $\begin{pmatrix} a & b \\ 3b & a \end{pmatrix}$ and $\begin{pmatrix} x & y \\ 3y & x \end{pmatrix}$. This proves the result that this bijection is an isomorphism.

4.4.5. Solution. As in the previous problem, it is very easy to see that these sets P_1 and P_2 are subrings of $\mathbb{Q}$.
No this map is not an isomorphism.

4.4.7. Solution. It is easy to see that K is a subring of $\mathbb{C}$.
Also f is a monomorphism from K into $\mathbb{C}$.

4.4.9. Solution. Since this quotient ring consist of three elements, the cosets $3\mathbb{Z}, 3\mathbb{Z}+1$, and $3\mathbb{Z}+2$, its order is 3. Therefore its additive group is a cyclic group of order 3.

4.4.11. Solution. Only in the case when $m = n$ or $m = -n$ do we actually get an isomorphism of rings.

4.4.13. Solution. We already know that there is a ring homomorphism $\theta : R \longrightarrow R/I$ defined by $\theta(r) = r+I$ and we now extend this by defining $\Theta : \mathbf{M}_2(R) \longrightarrow \mathbf{M}_2(R/I)$ by

$$\Theta\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = \begin{pmatrix} a+I & b+I \\ c+I & d+I \end{pmatrix}.$$

We need to check that Θ is a ring homomorphism.

Clearly Θ is an onto map. and $\mathbf{ker}\,\Theta = \mathbf{M}_2(I)$. The First Isomorphism Theorem now implies that $\mathbf{M}_2(R)/\mathbf{M}_2(I) \cong \mathbf{M}_2(R/I)$ as required.

4.4.15. Solution. First we note that J/I is an ideal of R/I, which can be easily verified. We define a function $\theta : R/I \longrightarrow R/J$ by $\theta(r+I) = r+J$, for all $r \in R$. The First Isomorphism Theorem now implies the result.

4.4.17. Solution. If R has an identity element then we take $K = R$. Therefore suppose that R does not have a multiplicative identity. It is enough to construct a monomorphism $f : R \longrightarrow K$ where K is the ring defined in Problem 4.4.16.

We define the mapping $f : R \longrightarrow K$ by $f(x) = (x, 0)$, where $x \in R$. Evidently, $\mathbf{Ker}f = \{0_R\}$ so f is a monomorphism. The result follows.

4.4.19. Solution. Let 1 be the identity element of R. Prove that $f(1)$ is the identity of S.

CHAPTER 5

5.1.1. Solution. First we remark that $P \neq \emptyset$. Let α, β be arbitrary elements of $\mathbb{Q}(\sqrt{2})$, $\alpha = a + b\sqrt{2}$, $\beta = a_1 + b_1\sqrt{2}$, where $a, b, a_1, b_1 \in \mathbb{Q}$. Then

$$\alpha - \beta = (a - a_1) + (b - b_1)\sqrt{2}, \alpha\beta = (aa_1 + 2bb_1) + (ab_1 + ba_1)\sqrt{2}.$$

It follows that $\alpha - \beta, \alpha\beta \in \mathbb{Q}(\sqrt{2})$. Clearly $1 \in \mathbb{Q}(\sqrt{2})$. Let $\alpha \neq 0$. Then it is easy to see that

$$\alpha^{-1} = \frac{a}{a^2 - 2b^2} + \left(\frac{-b}{a^2 - 2b^2}\right)\sqrt{2} \in \mathbb{Q}(\sqrt{2}).$$

Theorem 5.1.2 shows that $\mathbb{Q}(\sqrt{2})$ is a subfield of $\mathbb{C}$. More precisely, then, $\mathbb{Q}(\sqrt{2})$ is a subfield of $\mathbb{R}$.

5.1.3. Solution. There are no solutions to $x^2 - x - 3 = 0$ in $\mathbb{Q}(\sqrt{2})$.

5.1.5. Solution.

+	0	1	2	3	4	5	6
0	0	1	2	3	4	5	6
1	1	2	3	4	5	6	0
2	2	3	4	5	6	0	1
3	3	4	5	6	0	1	2
4	4	5	6	0	1	2	3
5	5	6	0	1	2	3	4
6	6	0	1	2	3	4	5

,

×	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

.

5.1.7. Solution. $x = 4$.

5.1.9. Solution. Suppose that f is an isomorphism from $\mathbb{Q}(\sqrt{5})$ onto $\mathbb{Q}(\sqrt{2})$. Show $f(a) = a$ for each $a \in \mathbb{Z}$. Suppose that $f(\sqrt{5}) = x + y\sqrt{2}$, for some $x, y \in \mathbb{Q}$. Then we have $5 = (x + y\sqrt{2})^2 = x^2 + 2y^2 + 2xy\sqrt{5}$. It follows that $2xy = 0$ since $\sqrt{5}$ is irrational. But then either $x = 0$ or $y = 0$ and we obtain a contradiction since $x^2 + 2y^2 = 5$ has no rational solutions.

5.1.11. Solution. Let also $g(X) = b_0 + b_1X + b_2X^2 + \cdots + b_nX^n$. (Note that we can use the same n for both $f(X)$ and $g(X)$ by allowing some of the b_i to be zero.)

Then $f(X)+g(X)=(a_0+b_0)+(a_1+b_1)X+(a_2+b_2)X^2+\ldots(a_i+b_i)X^i+\cdots+(a_n+b_n)X^n$. It follows that $D((f+g)(X))=(a_1+b_1)+2(a_2+b_2)X+3(a_3+b_3)X^2+\cdots+i(a_i+b_i)X^{i-1}+\cdots+n(a_n+b_n)X^{n-1}=(a_1+2a_2X+3a_3X^2+\cdots+na_nX^{n-1})+(b_1+2b_2X+3b_3X^2+\cdots+nb_nX^{n-1})=Df(X)+Dg(X)$.

5.1.13. Solution. Since f is an isomorphism we must have $f(1)=1$, because $f(1)=f(1^2)=f(1)^2$. It then follows that if n is a natural number we have $f(n)=f(n\cdot 1)=nf(1)=n$. Since $f(-n)=-f(n)$ the equation $f(n)=n$ also holds for all negative integers. Now we have $n=f(n)=f(m\cdot(n/m))=m\cdot f(n/m)$ and hence $f(n/m)=n/m$.

5.1.15. Solution. Form the ideal generated by X^2+2 which we call I and then construct $\mathbb{F}_5[X]/I$.

5.1.17. Solution. To construct a field with p^n elements we use Propositions 4.3.16 and 4.3.17. Work inside the ring $\mathbb{F}_p[X]$ and need a polynomial of degree n which is irreducible over this ring. Let $m(X)$ be such an irreducible polynomial. We then form the ideal generated by $m(X)$ which we call I and then construct $\mathbb{F}_p[X]/I$. This is the required field.

5.1.19. Solution. Of course if $\alpha(x+yi)=x+yi$ then we get the identity automorphism. But it is also easy to see that complex conjugation is an $\mathbb{R}$-automorphism, so there are two such automorphisms.

5.2.1. Solution. It is easy to see that $\mathbb{Q}(\sqrt{2}+\sqrt{3})$ is a subfield of $\mathbb{Q}(\sqrt{2},\sqrt{3})$. On the other hand $(\sqrt{2}+\sqrt{3})^2\in\mathbb{Q}(\sqrt{2}+\sqrt{3})$ also. Thus $\sqrt{6}\in\mathbb{Q}(\sqrt{2}+\sqrt{3})$. Hence $\sqrt{6}(\sqrt{2}+\sqrt{3})\in\mathbb{Q}(\sqrt{2}+\sqrt{3})$ also. It follows that $\sqrt{2}=2\sqrt{3}+3\sqrt{2}-2(\sqrt{2}+\sqrt{3})\in\mathbb{Q}(\sqrt{2}+\sqrt{3})$. Finally $\sqrt{3}=(\sqrt{2}+\sqrt{3})-\sqrt{2}\in\mathbb{Q}(\sqrt{2}+\sqrt{3})$ so $\mathbb{Q}(\sqrt{2},\sqrt{3})$ is a subset of $\mathbb{Q}(\sqrt{2}+\sqrt{3})$. The equality now follows.

5.2.3. Solution. $\mathbf{Q}(\sqrt{7})=\{a+b\sqrt{7}|a,b\in\mathbb{Q}\}$

5.2.5. Solution. As has been mentioned π is a transcendental number so the number π is not the root of any polynomial equation with rational coefficients. By the theory it follows that the elements of $\mathbb{Q}(\pi)$ are actually of the form $f(\pi)/g(\pi)$ where $f(X),g(X)$ are polynomials with rational coefficients.

5.2.7. Solution. The minimal polynomial of $\sqrt[3]{2}$ must be X^3-2 and the degree of $\mathbb{Q}(\sqrt[3]{2})$ over $\mathbb{Q}$ is 3. Then every element of $\mathbb{Q}(\sqrt[3]{2})$ has the form $a+b\sqrt[3]{2}+c\sqrt[3]{4}$. Since θ is an isomorphism it is a $\mathbb{Q}$-automorphism of $\mathbb{Q}(\sqrt[3]{2})$. In particular $\sqrt[3]{2}$ is mapped either to itself or to $\sqrt[3]{2}e^{2\pi i/3}$ or to $\sqrt[3]{2}e^{4\pi i/3}$. However the elements $\sqrt[3]{2}e^{2\pi i/3}$ and $\sqrt[3]{2}e^{4\pi i/3}$ are not real, whereas $\mathbb{Q}(\sqrt[3]{2})$ is a subfield of the real

numbers. Hence we must have $\theta(\sqrt[3]{2}) = \sqrt[3]{2}$ and since θ is a homomorphism it follows that θ must be the identity map.

5.2.9. Solution. $X^4 - 10X^2 + 1 = 0$.

5.3.1. Solution. A basis for $\mathbb{Q}(\sqrt{3})$ is $\{1, \sqrt{3}\}$ and its dimension over $\mathbb{Q}$ is 2.

5.3.3. Solution. Clearly $(1 + X + X^2 + \cdots + X^{p-1})(X - 1) = (X^p - 1)$. Hence $f(X) = (X^p - 1)/(X - 1)$. Consider $f(X + 1)$ and show each of the coefficients in its expansion is divisible by p, then use Eisenstein's Criterion.

5.3.5. Solution. $[\mathbb{Q}(\xi) : \mathbb{Q}] = 7$.
A basis will just be $\{1, \xi, \xi^2, \xi^3, \xi^4, \xi^5, \xi^6\}$.

5.3.7. Solution. The degree is 15.

5.3.9. Solution. $[\mathbb{Q}(\kappa) : \mathbb{Q}] = p$.

5.3.11. Solution. Suppose that $[K : F] = p$ and let $\alpha \in K \setminus F$. We can form the extension $F(\alpha)$ of F and use $[K : F] = [K : F(\alpha)][F(\alpha) : F]$.

5.3.13. Solution. We have F is a vector space over $\mathbb{F}_p$ with a basis with n elements. This makes a total of p^n elements.

INDEX

An Introduction to Essential Algebraic Structures, First Edition. Martyn R. Dixon, Leonid A. Kurdachenko and Igor Ya. Subbotin.

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.