

A BOOK OF ABSTRACT ALGEBRA

Second Edition

Charles C. Pinter
Professor of Mathematics
Bucknell University

Dover Publications, Inc., Mineola, New York

Copyright © 1982, 1990 by Charles C. Pinter
All rights reserved.

Bibliographical Note

This Dover edition, first published in 2010, is an unabridged republication of the 1990 second edition of the work originally published in 1982 by the McGraw-Hill Publishing Company, Inc., New York.

Library of Congress Cataloging-in-Publication Data

Pinter, Charles C, 1932–

A book of abstract algebra / Charles C. Pinter. — Dover ed.

p. cm.

Originally published: 2nd ed. New York : McGraw-Hill, 1990.

Includes bibliographical references and index.

ISBN-13: 978-0-486-47417-5

ISBN-10: 0-486-47417-8

1. Algebra, Abstract. I. Title.

QA162.P56 2010

512'.02—dc22

2009026228

Manufactured in the United States by Courier Corporation
47417803

www.doverpublications.com

Preface

Chapter 1 Why Abstract Algebra?

History of Algebra. New Algebras. Algebraic Structures. Axioms and Axiomatic Algebra. Abstraction in Algebra.

Chapter 2 Operations

Operations on a Set. Properties of Operations.

Chapter 3 The Definition of Groups

Groups. Examples of Infinite and Finite Groups. Examples of Abelian and Nonabelian Groups. Group Tables.

Theory of Coding: Maximum-Likelihood Decoding.

Chapter 4 Elementary Properties of Groups

Uniqueness of Identity and Inverses. Properties of Inverses.

Direct Product of Groups.

Chapter 5 Subgroups

Definition of Subgroup. Generators and Defining Relations.

Cayley Diagrams. Center of a Group. Group Codes; Hamming Code.

Chapter 6 Functions

Injective, Surjective, Bijective Function. Composite and Inverse of Functions.

Finite-State Machines. Automata and Their Semigroups.

Chapter 7 Groups of Permutations

Symmetric Groups. Dihedral Groups.

An Application of Groups to Anthropology.

Chapter 8 Permutations of a Finite Set

Decomposition of Permutations into Cycles. Transpositions. Even and Odd Permutations. Alternating Groups.

Chapter 9 Isomorphism

The Concept of Isomorphism in Mathematics. Isomorphic and Nonisomorphic Groups. Cayley's Theorem.

Chapter 10 Order of Group Elements

Powers/Multiples of Group Elements. Laws of Exponents. Properties of the Order of Group Elements.

Chapter 11 Cyclic Groups

Finite and Infinite Cyclic Groups. Isomorphism of Cyclic Groups. Subgroups of Cyclic Groups.

Chapter 12 Partitions and Equivalence Relations

Chapter 13 Counting Cosets

Lagrange's Theorem and Elementary Consequences.

Survey of Groups of Order ≤ 10 .

Number of Conjugate Elements. Group Acting on a Set.

Chapter 14 Homomorphisms

Elementary Properties of Homomorphisms. Normal Subgroups. Kernel and Range.

Inner Direct Products. Conjugate Subgroups.

Chapter 15 Quotient Groups

Quotient Group Construction. Examples and Applications.

The Class Equation. Induction on the Order of a Group.

Chapter 16 The Fundamental Homomorphism Theorem

Fundamental Homomorphism Theorem and Some Consequences.

The Isomorphism Theorems. The Correspondence Theorem. Cauchy's Theorem. Sylow Subgroups. Sylow's Theorem. Decomposition Theorem for Finite Abelian Groups.

Chapter 17 Rings: Definitions and Elementary Properties

Commutative Rings. Unity. Invertibles and Zero-Divisors. Integral Domain. Field.

Chapter 18 Ideals and Homomorphisms

Chapter 19 Quotient Rings

Construction of Quotient Rings. Examples. Fundamental Homomorphism Theorem and Some Consequences. Properties of Prime and Maximal Ideals.

Chapter 20 Integral Domains

Characteristic of an Integral Domain. Properties of the Characteristic. Finite Fields. Construction of the Field of Quotients.

Chapter 21 The Integers

Ordered Integral Domains. Well-ordering. Characterization of $\mathbf{Z}$ Up to Isomorphism. Mathematical Induction. Division Algorithm.

Chapter 22 Factoring into Primes

Ideals of $\mathbb{Z}$. Properties of the GCD. Relatively Prime Integers. Primes. Euclid's Lemma. Unique Factorization.

Chapter 23 Elements of Number Theory (Optional)

Properties of Congruence. Theorems of Fermat and Euler. Solutions of Linear Congruences. Chinese Remainder Theorem.

Wilson's Theorem and Consequences. Quadratic Residues. The Legendre Symbol. Primitive Roots.

Chapter 24 Rings of Polynomials

Motivation and Definitions. Domain of Polynomials over a Field. Division Algorithm.

Polynomials in Several Variables. Fields of Polynomial Quotients.

Chapter 25 Factoring Polynomials

Ideals of $F[x]$. Properties of the GCD. Irreducible Polynomials. Unique factorization.

Euclidean Algorithm.

Chapter 26 Substitution in Polynomials

Roots and Factors. Polynomial Functions. Polynomials over $\mathbb{Q}$. Eisenstein's Irreducibility Criterion. Polynomials over the Reals. Polynomial Interpolation.

Chapter 27 Extensions of Fields

Algebraic and Transcendental Elements. The Minimum Polynomial. Basic Theorem on Field Extensions.

Chapter 28 Vector Spaces

Elementary Properties of Vector Spaces. Linear Independence. Basis. Dimension. Linear Transformations.

Chapter 29 Degrees of Field Extensions

Simple and Iterated Extensions. Degree of an Iterated Extension.

Fields of Algebraic Elements. Algebraic Numbers. Algebraic Closure.

Chapter 30 Ruler and Compass

Constructible Points and Numbers. Impossible Constructions.

Constructible Angles and Polygons.

Chapter 31 Galois Theory: Preamble

Multiple Roots. Root Field. Extension of a Field. Isomorphism.

Roots of Unity. Separable Polynomials. Normal Extensions.

Chapter 32 Galois Theory: The Heart of the Matter

Field Automorphisms. The Galois Group. The Galois Correspondence. Fundamental Theorem of Galois Theory.

Computing Galois Groups.

Chapter 33 Solving Equations by Radicals

Radical Extensions. Abelian Extensions. Solvable Groups. Insolubility of the Quintic.

Appendix A	Review of Set Theory
Appendix B	Review of the Integers
Appendix C	Review of Mathematical Induction
	Answers to Selected Exercises
	Index

* Italic headings indicate topics discussed in the exercise sections.

WHY ABSTRACT ALGEBRA?

When we open a textbook of abstract algebra for the first time and peruse the table of contents, we are struck by the unfamiliarity of almost every topic we see listed. Algebra is a subject we know well, but here it looks surprisingly different. What are these differences, and how fundamental are they?

First, there is a major difference in emphasis. In elementary algebra we learned the basic symbolism and methodology of algebra; we came to see how problems of the real world can be reduced to sets of equations and how these equations can be solved to yield numerical answers. This technique for translating complicated problems into symbols is the basis for all further work in mathematics and the exact sciences, and is one of the triumphs of the human mind. However, algebra is not only a technique, it is also a branch of learning, a *discipline*, like calculus or physics or chemistry. It is a coherent and unified body of knowledge which may be studied systematically, starting from first principles and building up. So the first difference between the elementary and the more advanced course in algebra is that, whereas earlier we concentrated on technique, we will now develop that branch of mathematics called algebra in a systematic way. Ideas and general principles will take precedence over problem solving. (By the way, this does not mean that modern algebra has no applications—quite the opposite is true, as we will see soon.)

Algebra at the more advanced level is often described as *modern* or *abstract* algebra. In fact, both of these descriptions are partly misleading. Some of the great discoveries in the upper reaches of present-day algebra (for example, the so-called Galois theory) were known many years before the American Civil War; and the broad aims of algebra today were clearly stated by Leibniz in the seventeenth century. Thus, “modern” algebra is not so very modern, after all! To what extent is it *abstract*? Well, abstraction is all relative; one person’s abstraction is another person’s bread and butter. The abstract tendency in mathematics is a little like the situation of changing moral codes, or changing tastes in music: What shocks one generation becomes the norm in the next. This has been true throughout the history of mathematics.

For example, 1000 years ago negative numbers were considered to be an outrageous idea. After all, it was said, numbers are for counting: we may have one orange, or two oranges, or no oranges at all; but how can we have minus an orange? The *logisticians*, or professional calculators, of those days used negative numbers as an aid in their computations; they considered these numbers to be a useful fiction, for if you believe in them then every linear equation $ax + b = 0$ has a solution (namely $x = -b/a$, provided $a \neq 0$). Even the great Diophantus once described the solution of $4x + 6 = 2$ as an *absurd* number. The idea of a system of numeration which included negative numbers was far too abstract for many of the learned heads of the tenth century!

The history of the complex numbers (numbers which involve $\sqrt{-1}$) is very much the same. For hundreds of years, mathematicians refused to accept them because they couldn't find concrete examples or applications. (They are now a basic tool of physics.)

Set theory was considered to be highly abstract a few years ago, and so were other commonplaces of today. Many of the abstractions of modern algebra are already being used by scientists, engineers, and computer specialists in their everyday work. They will soon be common fare, respectably “concrete,” and by then there will be new “abstractions.”

Later in this chapter we will take a closer look at the particular brand of abstraction used in algebra. We will consider how it came about and why it is useful.

Algebra has evolved considerably, especially during the past 100 years. Its growth has been closely linked with the development of other branches of mathematics, and it has been deeply influenced by philosophical ideas on the nature of mathematics and the role of logic. To help us understand the nature and spirit of modern algebra, we should take a brief look at its origins.

ORIGINS

The order in which subjects follow each other in our mathematical education tends to repeat the historical stages in the evolution of mathematics. In this scheme, elementary algebra corresponds to the great classical age of algebra, which spans about 300 years from the sixteenth through the eighteenth centuries. It was during these years that the art of solving equations became highly developed and modern symbolism was invented.

The word “algebra”—*al jebr* in Arabic—was first used by Mohammed of Kharizm, who taught mathematics in Baghdad during the ninth century. The word may be roughly translated as “reunion,” and describes his method for collecting the terms of an equation in order to solve it. It is an amusing fact that the word “algebra” was first used in Europe in quite another context. In Spain barbers were called *algebristas*, or bonesetters (they *reunited* broken bones), because medieval barbers did bonesetting and bloodletting as a sideline to their usual business.

The origin of the word clearly reflects the actual context of algebra at that time, for it was mainly concerned with ways of solving equations. In fact, Omar Khayyam, who is best remembered for his brilliant verses on wine, song, love, and friendship which are collected in the *Rubaiyat*—but who was also a great mathematician—explicitly defined algebra as the *science of solving equations*.

Thus, as we enter upon the threshold of the classical age of algebra, its central theme is clearly identified as that of solving equations. Methods of solving the linear equation $ax + b = 0$ and the quadratic $ax^2 + bx + c = 0$ were well known even before the Greeks. But nobody had yet found a general solution for *cubic* equations

$$x^3 + ax^2 + bx = c$$

or *quartic* (fourth-degree) equations

$$x^4 + ax^3 + bx^2 + cx = d$$

This great accomplishment was the triumph of sixteenth century algebra.

The setting is Italy and the time is the Renaissance—an age of high adventure and brilliant achievement, when the wide world was reawakening after the long austerity of the Middle Ages. America had just been discovered, classical knowledge had been brought to light, and prosperity had

returned to the great cities of Europe. It was a heady age when nothing seemed impossible and even the old barriers of birth and rank could be overcome. Courageous individuals set out for great adventures in the far corners of the earth, while others, now confident once again of the power of the human mind, were boldly exploring the limits of knowledge in the sciences and the arts. The ideal was to be bold and many-faceted, to “know something of everything, and everything of at least one thing.” The great traders were patrons of the arts, the finest minds in science were adepts at political intrigue and high finance. The study of algebra was reborn in this lively milieu.

Those men who brought algebra to a high level of perfection at the beginning of its classical age—all typical products of the Italian Renaissance—were as colorful and extraordinary a lot as have ever appeared in a chapter of history. Arrogant and unscrupulous, brilliant, flamboyant, swaggering, and remarkable, they lived their lives as they did their work: with style and panache, in brilliant dashes and inspired leaps of the imagination.

The spirit of scholarship was not exactly as it is today. These men, instead of publishing their discoveries, kept them as well-guarded secrets to be used against each other in problem-solving competitions. Such contests were a popular attraction: heavy bets were made on the rival parties, and their reputations (as well as a substantial purse) depended on the outcome.

One of the most remarkable of these men was Girolamo Cardan. Cardan was born in 1501 as the illegitimate son of a famous jurist of the city of Pavia. A man of passionate contrasts, he was destined to become famous as a physician, astrologer, and mathematician—and notorious as a compulsive gambler, scoundrel, and heretic. After he graduated in medicine, his efforts to build up a medical practice were so unsuccessful that he and his wife were forced to seek refuge in the poorhouse. With the help of friends he became a lecturer in mathematics, and, after he cured the child of a senator from Milan, his medical career also picked up. He was finally admitted to the college of physicians and soon became its rector. A brilliant doctor, he gave the first clinical description of typhus fever, and as his fame spread he became the personal physician of many of the high and mighty of his day.

Cardan's early interest in mathematics was not without a practical side. As an inveterate gambler he was fascinated by what he recognized to be the laws of chance. He wrote a gamblers' manual entitled *Book on Games of Chance*, which presents the first systematic computations of probabilities. He also needed mathematics as a tool in casting horoscopes, for his fame as an astrologer was great and his predictions were highly regarded and sought after. His most important achievement was the publication of a book called *Ars Magna (The Great Art)*, in which he presented systematically all the algebraic knowledge of his time. However, as already stated, much of this knowledge was the personal secret of its practitioners, and had to be wheedled out of them by cunning and deceit. The most important accomplishment of the day, the general solution of the cubic equation which had been discovered by Tartaglia, was obtained in that fashion.

Tartaglia's life was as turbulent as any in those days. Born with the name of Niccolo Fontana about 1500, he was present at the occupation of Brescia by the French in 1512. He and his father fled with many others into a cathedral for sanctuary, but in the heat of battle the soldiers massacred the hapless citizens even in that holy place. The father was killed, and the boy, with a split skull and a deep saber cut across his jaws and palate, was left for dead. At night his mother stole into the cathedral and managed to carry him off; miraculously he survived. The horror of what he had witnessed caused him to stammer for the rest of his life, earning him the nickname *Tartaglia*, “the stammerer,” which he eventually adopted.

Tartaglia received no formal schooling, for that was a privilege of rank and wealth. However, he taught himself mathematics and became one of the most gifted mathematicians of his day. He translated Euclid and Archimedes and may be said to have originated the science of ballistics, for he wrote a

treatise on gunnery which was a pioneering effort on the laws of falling bodies.

In 1535 Tartaglia found a way of solving any cubic equation of the form $x^3 + ax^2 = b$ (that is, without an x term). When he announced his accomplishment (without giving any details, of course), he was challenged to an algebra contest by a certain Antonio Fior, a pupil of the celebrated professor of mathematics Scipio del Ferro. Scipio had already found a method for solving any cubic equation of the form $x^3 + ax = b$ (that is, without an x^2 term), and had confided his secret to his pupil Fior. It was agreed that each contestant was to draw up 30 problems and hand the list to his opponent. Whoever solved the greater number of problems would receive a sum of money deposited with a lawyer. A few days before the contest, Tartaglia found a way of extending his method so as to solve *any* cubic equation. In less than 2 hours he solved all his opponent's problems, while his opponent failed to solve even one of those proposed by Tartaglia.

For some time Tartaglia kept his method for solving cubic equations to himself, but in the end he succumbed to Cardan's accomplished powers of persuasion. Influenced by Cardan's promise to help him become artillery adviser to the Spanish army, he revealed the details of his method to Cardan under the promise of strict secrecy. A few years later, to Tartaglia's unbelieving amazement and indignation, Cardan published Tartaglia's method in his book *Ars Magna*. Even though he gave Tartaglia full credit as the originator of the method, there can be no doubt that he broke his solemn promise. A bitter dispute arose between the mathematicians, from which Tartaglia was perhaps lucky to escape alive. He lost his position as public lecturer at Brescia, and lived out his remaining years in obscurity.

The next great step in the progress of algebra was made by another member of the same circle. It was Ludovico Ferrari who discovered the general method for solving quartic equations—equations of the form

$$x^4 + ax^3 + bx^2 + cx = d$$

Ferrari was Cardan's personal servant. As a boy in Cardan's service he learned Latin, Greek, and mathematics. He won fame after defeating Tartaglia in a contest in 1548, and received an appointment as supervisor of tax assessments in Mantua. This position brought him wealth and influence, but he was not able to dominate his own violent disposition. He quarreled with the regent of Mantua, lost his position, and died at the age of 43. Tradition has it that he was poisoned by his sister.

As for Cardan, after a long career of brilliant and unscrupulous achievement, his luck finally abandoned him. Cardan's son poisoned his unfaithful wife and was executed in 1560. Ten years later, Cardan was arrested for heresy because he published a horoscope of Christ's life. He spent several months in jail and was released after renouncing his heresy privately, but lost his university position and the right to publish books. He was left with a small pension which had been granted to him, for some unaccountable reason, by the Pope.

As this colorful time draws to a close, algebra emerges as a major branch of mathematics. It became clear that methods can be found to solve many different types of equations. In particular, formulas had been discovered which yielded the roots of all cubic and quartic equations. Now the challenge was clearly out to take the next step, namely, to find a formula for the roots of equations of degree 5 or higher (in other words, equations with an x^5 term, or an x^6 term, or higher). During the next 200 years, there was hardly a mathematician of distinction who did not try to solve this problem, but none succeeded. Progress was made in new parts of algebra, and algebra was linked to geometry with the invention of analytic geometry. But the problem of solving equations of degree higher than 4 remained unsettled. It was, in the expression of Lagrange, "a challenge to the human mind."

It was therefore a great surprise to all mathematicians when in 1824 the work of a young

Norwegian prodigy named Niels Abel came to light. In his work, Abel showed that *there does not exist any formula* (in the conventional sense we have in mind) for the roots of an algebraic equation whose degree is 5 or greater. This sensational discovery brings to a close what is called the classical age of algebra. Throughout this age algebra was conceived essentially as the science of solving equations, and now the outer limits of this quest had apparently been reached. In the years ahead, algebra was to strike out in new directions.

THE MODERN AGE

About the time Niels Abel made his remarkable discovery, several mathematicians, working independently in different parts of Europe, began raising questions about algebra which had never been considered before. Their researches in different branches of mathematics had led them to investigate “algebras” of a very unconventional kind—and in connection with these algebras they had to find answers to questions which had nothing to do with solving equations. Their work had important applications, and was soon to compel mathematicians to greatly enlarge their conception of what algebra is about.

The new varieties of algebra arose as a perfectly natural development in connection with the application of mathematics to practical problems. This is certainly true for the example we are about to look at first.

The Algebra of Matrices

A *matrix* is a rectangular array of numbers such as

$$\begin{pmatrix} 2 & 11 & -3 \\ 9 & 0.5 & 4 \end{pmatrix}$$

Such arrays come up naturally in many situations, for example, in the solution of simultaneous linear equations. The above matrix, for instance, is the *matrix of coefficients* of the pair of equations

$$2x + 11y - 3z = 0$$

$$9x + 0.5y + 4z = 0$$

Since the solution of this pair of equations depends only on the coefficients, we may solve it by working on the matrix of coefficients alone and ignoring everything else.

We may consider the entries of a matrix to be arranged in *rows* and *columns*; the above matrix has two rows which are

$$(2 \ 11 \ -3) \quad \text{and} \quad (9 \ 0.5 \ 4)$$

and three columns which are

$$\begin{pmatrix} 2 \\ 9 \end{pmatrix} \quad \begin{pmatrix} 11 \\ 0.5 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -3 \\ 4 \end{pmatrix}$$

It is a 2×3 matrix.

To simplify our discussion, we will consider only 2×2 matrices in the remainder of this section.

Matrices are added by adding corresponding entries:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} a+a' & b+b' \\ c+c' & d+d' \end{pmatrix}$$

The matrix

$$\mathbf{0} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

is called the *zero matrix* and behaves, under addition, like the number zero.

The multiplication of matrices is a little more difficult. First, let us recall that the *dot product* of two vectors (a, b) and (a', b') is

$$(a, b) \cdot (a', b') = aa' + bb'$$

that is, we multiply corresponding components and add. Now, suppose we want to multiply two matrices $\mathbf{A}$ and $\mathbf{B}$; we obtain the product $\mathbf{AB}$ as follows:

The entry in the first row and first column of $\mathbf{AB}$, that is, in *this* position

$$\begin{pmatrix} x & | & - & - \\ - & | & - & - \\ - & | & - & - \\ - & | & - & - \end{pmatrix}$$

is equal to the dot product of the first row of $\mathbf{A}$ by the first column of $\mathbf{B}$. The entry in the first row and *second* column of $\mathbf{AB}$, in other words, *this* position

$$\begin{pmatrix} - & | & - & x \\ - & | & - & - \\ - & | & - & - \\ - & | & - & - \end{pmatrix}$$

is equal to the dot product of the first row of $\mathbf{A}$ by the *second* column of $\mathbf{B}$. And so on. For example,

$$\begin{pmatrix} \boxed{1} & \boxed{2} \\ 3 & 0 \end{pmatrix} \begin{pmatrix} \boxed{1} & 1 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 5 & \\ & \end{pmatrix}$$

$$\begin{pmatrix} \boxed{1} & \boxed{2} \\ 3 & 0 \end{pmatrix} \begin{pmatrix} 1 & \boxed{1} \\ 2 & \boxed{0} \end{pmatrix} = \begin{pmatrix} & 1 \\ & \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ \boxed{3} & \boxed{0} \end{pmatrix} \begin{pmatrix} \boxed{1} & 1 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} & \\ 3 & \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 \\ \boxed{3} & \boxed{0} \end{pmatrix} \begin{pmatrix} 1 & \boxed{1} \\ 2 & \boxed{0} \end{pmatrix} = \begin{pmatrix} & \\ & 3 \end{pmatrix}$$

So finally,

$$\begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 5 & 1 \\ 3 & 3 \end{pmatrix}$$

The rules of algebra for matrices are very different from the rules of “conventional” algebra. For instance, the commutative law of multiplication, $\mathbf{AB} = \mathbf{BA}$, is not true. Here is a simple example:

$$\underbrace{\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}}_{\mathbf{B}} = \underbrace{\begin{pmatrix} 2 & 1 \\ 2 & 1 \end{pmatrix}}_{\mathbf{AB}} \neq \underbrace{\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}}_{\mathbf{BA}} = \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}}_{\mathbf{B}} \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}}_{\mathbf{A}}$$

If A is a real number and $A^2 = 0$, then necessarily $A = 0$; but this is not true of matrices. For

example,

$$\underbrace{\begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}}_{\mathbf{A}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

that is, $\mathbf{A}^2 = \mathbf{0}$ although $\mathbf{A} \neq \mathbf{0}$.

In the algebra of numbers, if $AB = AC$ where $A \neq 0$, we may cancel A and conclude that $B = C$. In matrix algebra we cannot. For example,

$$\underbrace{\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}}_{\mathbf{B}} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}}_{\mathbf{C}}$$

that is, $\mathbf{AB} = \mathbf{AC}$, $\mathbf{A} \neq \mathbf{0}$, yet $\mathbf{B} \neq \mathbf{C}$.

The *identity* matrix

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

corresponds in matrix multiplication to the number 1; for we have $\mathbf{AI} = \mathbf{IA} = \mathbf{A}$ for every 2×2 matrix $\mathbf{A}$. If A is a number and $A^2 = 1$, we conclude that $A = \pm 1$. Matrices do not obey this rule. For example,

$$\underbrace{\begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & -1 \end{pmatrix}}_{\mathbf{A}} = \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\mathbf{I}}$$

that is, $\mathbf{A}^2 = \mathbf{I}$, and yet $\mathbf{A}$ is neither $\mathbf{I}$ nor $-\mathbf{I}$.

No more will be said about the algebra of matrices at this point, except that we must be aware, once again, that it is a new game whose rules are quite different from those we apply in conventional algebra.

Boolean Algebra

An even more bizarre kind of algebra was developed in the mid-nineteenth century by an Englishman named George Boole. This algebra—subsequently named boolean algebra after its inventor—has a myriad of applications today. It is formally the same as the algebra of sets.

If S is a set, we may consider *union* and *intersection* to be operations on the subsets of S . Let us agree provisionally to write

$$A + B \quad \text{for} \quad A \cup B$$

and

$$A \cdot B \quad \text{for} \quad A \cap B$$

(This convention is not unusual.) Then,

$$A + B = B + A \qquad A \cdot B = B \cdot A$$

$$A \cdot (B + C) = A \cdot B + A \cdot C$$

$$A + \emptyset = A \qquad A \cdot \emptyset = \emptyset$$

and so on.

These identities are analogous to the ones we use in elementary algebra. But the following identities are also true, and they have no counterpart in conventional algebra:

$$A + (B \cdot C) = (A + B) \cdot (A + C)$$

$$A + A = A \qquad A \cdot A = A$$

$$(A + B) \cdot A = A \qquad (A \cdot B) + A = A$$

and so on.

This unusual algebra has become a familiar tool for people who work with electrical networks, computer systems, codes, and so on. It is as different from the algebra of numbers as it is from the algebra of matrices.

Other exotic algebras arose in a variety of contexts, often in connection with scientific problems. There were “complex” and “hypercomplex” algebras, algebras of vectors and tensors, and many others. Today it is estimated that over 200 different kinds of algebraic systems have been studied, each of which arose in connection with some application or specific need.

Algebraic Structures

As legions of new algebras began to occupy the attention of mathematicians, the awareness grew that algebra can no longer be conceived merely as the *science of solving equations*. It had to be viewed much more broadly as a branch of mathematics capable of revealing general principles which apply equally to *all known and all possible algebras*.

What is it that all algebras have in common? What trait do they share which lets us refer to all of them as “algebras”? In the most general sense, every algebra consists of a *set* (a set of numbers, a set of matrices, a set of switching components, or any other kind of set) and certain *operations* on that set. An operation is simply a way of combining any two members of a set to produce a unique third member of the same set.

Thus, we are led to the modern notion of algebraic structure. An *algebraic structure* is understood to be an arbitrary set, with one or more operations defined on it. And algebra, then, is defined to be *the study of algebraic structures*.

It is important that we be awakened to the full generality of the notion of algebraic structure. We must make an effort to discard all our preconceived notions of what an algebra is, and look at this new notion of algebraic structure in its naked simplicity. *Any* set, with a rule (or rules) for combining its elements, is already an algebraic structure. There does not need to be any connection with known mathematics. For example, consider the set of all colors (pure colors as well as color combinations), and the operation of mixing any two colors to produce a new color. This may be conceived as an algebraic structure. It obeys certain rules, such as the commutative law (mixing red and blue is the same as mixing blue and red). In a similar vein, consider the set of all musical sounds with the operation of combining any two sounds to produce a new (harmonious or disharmonious) combination.

As another example, imagine that the guests at a family reunion have made up a rule for picking

the *closest common relative* of any two persons present at the reunion (and suppose that, for any two people at the reunion, their closest common relative is also present at the reunion). This too, is an algebraic structure: we have a set (namely the set of persons at the reunion) and an operation on that set (namely the “closest common relative” operation).

As the general notion of algebraic structure became more familiar (it was not fully accepted until the early part of the twentieth century), it was bound to have a profound influence on what mathematicians perceived algebra to *be*. In the end it became clear that the purpose of algebra is to study algebraic structures, and nothing less than that. Ideally it should aim to be a general science of algebraic structures whose results should have applications to particular cases, thereby making contact with the older parts of algebra. Before we take a closer look at this program, we must briefly examine another aspect of modern mathematics, namely, the increasing use of the axiomatic method.

AXIOMS

The axiomatic method is beyond doubt the most remarkable invention of antiquity, and in a sense the most puzzling. It appeared suddenly in Greek geometry in a highly developed form—already sophisticated, elegant, and thoroughly modern in style. Nothing seems to have foreshadowed it and it was unknown to ancient mathematicians before the Greeks. It appears for the first time in the light of history in the great textbook of early geometry, Euclid’s *Elements*. Its origins—the first tentative experiments in formal deductive reasoning which must have preceded it—remain steeped in mystery.

Euclid’s *Elements* embodies the axiomatic method in its purest form. This amazing book contains 465 geometric propositions, some fairly simple, some of astounding complexity. What is really remarkable, though, is that the 465 propositions, forming the largest body of scientific knowledge in the ancient world, are derived logically from only 10 premises which would pass as trivial observations of common sense. Typical of the premises are the following:

Things equal to the same thing are equal to each other.

The whole is greater than the part.

A straight line can be drawn through any two points.

All right angles are equal.

So great was the impression made by Euclid’s *Elements* on following generations that it became the model of correct mathematical form and remains so to this day.

It would be wrong to believe there was no notion of demonstrative mathematics before the time of Euclid. There is evidence that the earliest geometers of the ancient Middle East used reasoning to discover geometric principles. They found proofs and must have hit upon many of the same proofs we find in Euclid. The difference is that Egyptian and Babylonian mathematicians considered logical demonstration to be an auxiliary process, like the preliminary sketch made by artists—a private mental process which guided them to a result but did not deserve to be recorded. Such an attitude shows little understanding of the true nature of geometry and does not contain the seeds of the axiomatic method.

It is also known today that many—maybe most—of the geometric theorems in Euclid’s *Elements* came from more ancient times, and were probably borrowed by Euclid from Egyptian and Babylonian sources. However, this does not detract from the greatness of his work. Important as are the contents of the *Elements*, what has proved far more important for posterity is the formal manner in which Euclid presented these contents. The heart of the matter was the way he *organized* geometric facts—arranged them into a logical sequence where each theorem builds on preceding theorems and then forms the

logical basis for other theorems.

(We must carefully note that the axiomatic method is not a way of discovering facts but of organizing them. New facts in mathematics are found, as often as not, by inspired guesses or experienced intuition. To be accepted, however, they should be supported by proof in an axiomatic system.)

Euclid's *Elements* has stood throughout the ages as the model of organized, rational thought carried to its ultimate perfection. Mathematicians and philosophers in every generation have tried to imitate its lucid perfection and flawless simplicity. Descartes and Leibniz dreamed of organizing all human knowledge into an axiomatic system, and Spinoza created a deductive system of ethics patterned after Euclid's geometry. While many of these dreams have proved to be impractical, the method popularized by Euclid has become the prototype of modern mathematical form. Since the middle of the nineteenth century, the axiomatic method has been accepted as the only correct way of organizing mathematical knowledge.

To perceive why the axiomatic method is truly central to mathematics, we must keep one thing in mind: mathematics by its nature is essentially *abstract*. For example, in geometry straight lines are not stretched threads, but a concept obtained by disregarding all the properties of stretched threads except that of extending in one direction. Similarly, the concept of a geometric figure is the result of idealizing from all the properties of actual objects and retaining only their spatial relationships. Now, since the objects of mathematics are *abstractions*, it stands to reason that we must acquire knowledge about them by logic and not by observation or experiment (for how can one experiment with an abstract thought?).

This remark applies very aptly to modern algebra. The notion of algebraic structure is obtained by idealizing from all particular, concrete systems of algebra. We choose to ignore the properties of the actual objects in a system of algebra (they may be numbers, or matrices, or whatever—we disregard what they *are*), and we turn our attention simply to the way they combine under the given operations. In fact, just as we disregard what the objects in a system *are*, we also disregard what the operations *do* to them. We retain only the equations and inequalities which hold in the system, for only these are relevant to algebra. Everything else may be discarded. Finally, equations and inequalities may be deduced from one another logically, just as spatial relationships are deduced from each other in geometry.

THE AXIOMATICS OF ALGEBRA

Let us remember that in the mid-nineteenth century, when eccentric new algebras seemed to show up at every turn in mathematical research, it was finally understood that sacrosanct laws such as the identities $ab = ba$ and $a(bc) = (ab)c$ are not inviolable—for there are algebras in which they do not hold. By varying or deleting some of these identities, or by replacing them by new ones, an enormous variety of new systems can be created.

Most importantly, mathematicians slowly discovered that all the algebraic laws which hold in any system can be derived from a few simple, basic ones. This is a genuinely remarkable fact, for it parallels the discovery made by Euclid that a few very simple geometric postulates are sufficient to prove all the theorems of geometry. As it turns out, then, we have the same phenomenon in algebra: a few simple algebraic equations offer themselves naturally as axioms, and from them all other facts may be proved.

These basic algebraic laws are familiar to most high school students today. We list them here for reference. We assume that A is any set and there is an operation on A which we designate with the symbol $*$

$$a * b = b * a \quad (1)$$

If [Equation \(1\)](#) is true for any two elements a and b in A , we say that the operation $*$ is *commutative*.

What it means, of course, is that the value of $a * b$ (or $b * a$) is independent of the order in which a and b are taken.

$$a * (b * c) = (a * b) * c \quad (2)$$

If Equation (2) is true for any three elements a , b , and c in A , we say the operation $*$ is *associative*. Remember that an operation is a rule for combining any *two* elements, so if we want to combine *three* elements, we can do so in different ways. If we want to combine a , b , and c *without changing their order*, we may either combine a with the result of combining b and c , which produces $a * (b * c)$; or we may first combine a with b , and then combine the result with c , producing $(a * b) * c$. The associative law asserts that these two possible ways of combining three elements (without changing their order) yield the same result.

There exists an element e in A such that

$$e * a = a \quad \text{and} \quad a * e = a \quad \text{for every } a \text{ in } A \quad (3)$$

If such an element e exists in A , we call it an *identity element* for the operation $*$. An identity element is sometimes called a “neutral” element, for it may be combined with any element a without altering a . For example, 0 is an identity element for addition, and 1 is an identity element for multiplication.

For every element a in A , there is an element a^{-1} (“ a inverse”) in A such that

$$a * a^{-1} = e \quad \text{and} \quad a^{-1} * a = e \quad (4)$$

If statement (4) is true in a system of algebra, we say that every element has an inverse with respect to the operation $*$. The meaning of the inverse should be clear: the combination of any element with its inverse produces the neutral element (one might roughly say that the inverse of a “neutralizes” a). For example, if A is a set of numbers and the operation is addition, then the inverse of any number a is $(-a)$; if the operation is multiplication, the inverse of any $a \neq 0$ is $1/a$.

Let us assume now that the same set A has a second operation, symbolized by $\perp$, as well as the operation $*$:

$$a * (b \perp c) = (a * b) \perp (a * c) \quad (5)$$

If Equation (5) holds for any three elements a , b , and c in A , we say that $*$ is *distributive* over $\perp$. If there are two operations in a system, they must interact in some way; otherwise there would be no need to consider them together. The distributive law is the most common way (but not the only possible one) for two operations to be related to one another.

There are other “basic” laws besides the five we have just seen, but these are the most common ones. The most important algebraic systems have axioms chosen from among them. For example, when a mathematician nowadays speaks of a *ring*, the mathematician is referring to a set A with two operations, usually symbolized by $+$ and $\cdot$, having the following axioms:

Addition is commutative and associative, it has a neutral element commonly symbolized by 0, and every element a has an inverse $-a$ with respect to addition. Multiplication is associative, has a neutral element 1, and is distributive over addition.

Matrix algebra is a particular example of a ring, and all the laws of matrix algebra may be proved from the preceding axioms. However, there are many other examples of rings: rings of numbers, rings of functions, rings of code “words,” rings of switching components, and a great many more. Every algebraic law which can be proved in a ring (from the preceding axioms) is true in every *example* of a ring. In other words, instead of proving the same formula repeatedly—once for numbers, once for matrices, once for switching components, and so on—it is sufficient nowadays to prove only that the formula holds in rings, and then of necessity it will be true in all the hundreds of different concrete examples of rings.

By varying the possible choices of axioms, we can keep creating new axiomatic systems of algebra endlessly. We may well ask: is it legitimate to study *any* axiomatic system, with *any* choice of axioms, regardless of usefulness, relevance, or applicability? There are “radicals” in mathematics who claim the freedom for mathematicians to study any system they wish, without the need to justify it. However, the practice in established mathematics is more conservative: particular axiomatic systems are investigated on account of their relevance to new and traditional problems and other parts of mathematics, or because they correspond to particular applications.

In practice, how is a particular choice of algebraic axioms made? Very simply: when mathematicians look at different parts of algebra and notice that a common pattern of proofs keeps recurring, and essentially the same assumptions need to be made each time, they find it natural to single out this choice of assumptions as the axioms for a new system. All the important new systems of algebra were created in this fashion.

ABSTRACTION REVISITED

Another important aspect of axiomatic mathematics is this: when we capture mathematical facts in an axiomatic system, we never try to reproduce the facts in full, but only that side of them which is important or relevant in a particular context. This process of *selecting what is relevant* and disregarding everything else is the very essence of abstraction.

This kind of abstraction is so natural to us as human beings that we practice it all the time without being aware of doing so. Like the Bourgeois Gentleman in Molière’s play who was amazed to learn that he spoke in prose, some of us may be surprised to discover how much we think in abstractions. Nature presents us with a myriad of interwoven facts and sensations, and we are challenged at every instant to single out those which are immediately relevant and discard the rest. In order to make our surroundings comprehensible, we must continually pick out certain data and separate them from everything else.

For natural scientists, this process is the very core and essence of what they do. Nature is not made up of forces, velocities, and moments of inertia. Nature is a whole—nature simply *is*! The physicist isolates certain aspects of nature from the rest and finds the laws which govern these *abstractions*.

It is the same with mathematics. For example, the system of the integers (whole numbers), as known by our intuition, is a complex reality with many facets. The mathematician separates these facets from one another and studies them individually. From one point of view the set of the integers, with addition and multiplication, forms a *ring* (that is, it satisfies the axioms stated previously). From another point of view it is an ordered set, and satisfies special axioms of ordering. On a different level, the positive integers form the basis of “recursion theory,” which singles out the particular way positive integers may be *constructed*, beginning with 1 and adding 1 each time.

It therefore happens that the traditional subdivision of mathematics into subject matters has been radically altered. No longer are the integers one subject, complex numbers another, matrices another, and so on; instead, particular *aspects* of these systems are isolated, put in axiomatic form, and studied abstractly without reference to any specific objects. The other side of the coin is that each aspect is shared by many of the traditional systems: for example, algebraically the integers form a ring, and so do the complex numbers, matrices, and many other kinds of objects.

There is nothing intrinsically new about this process of divorcing properties from the actual objects *having* the properties; as we have seen, it is precisely what geometry has done for more than 2000 years. Somehow, it took longer for this process to take hold in algebra.

The movement toward axiomatics and abstraction in modern algebra began about the 1830s and was completed 100 years later. The movement was tentative at first, not quite conscious of its aims, but

it gained momentum as it converged with similar trends in other parts of mathematics. The thinking of many great mathematicians played a decisive role, but none left a deeper or longer lasting impression than a very young Frenchman by the name of Évariste Galois.

The story of Évariste Galois is probably the most fantastic and tragic in the history of mathematics. A sensitive and prodigiously gifted young man, he was killed in a duel at the age of 20, ending a life which in its brief span had offered him nothing but tragedy and frustration. When he was only a youth his father committed suicide, and Galois was left to fend for himself in the labyrinthine world of French university life and student politics. He was twice refused admittance to the École Polytechnique, the most prestigious scientific establishment of its day, probably because his answers to the entrance examination were too original and unorthodox. When he presented an early version of his important discoveries in algebra to the great academician Cauchy, this gentleman did not read the young student's paper, but lost it. Later, Galois gave his results to Fourier in the hope of winning the mathematics prize of the Academy of Sciences. But Fourier died, and that paper, too, was lost. Another paper submitted to Poisson was eventually returned because Poisson did not have the interest to read it through.

Galois finally gained admittance to the École Normale, another focal point of research in mathematics, but he was soon expelled for writing an essay which attacked the king. He was jailed twice for political agitation in the student world of Paris. In the midst of such a turbulent life, it is hard to believe that Galois found time to create his colossally original theories on algebra.

What Galois did was to tie in the problem of finding the roots of equations with new discoveries on groups of permutations. He explained exactly *which* equations of degree 5 or higher have solutions of the traditional kind—and which others do not. Along the way, he introduced some amazingly original and powerful concepts, which form the framework of much algebraic thinking to this day. Although Galois did not work explicitly in axiomatic algebra (which was unknown in his day), the abstract notion of algebraic structure is clearly prefigured in his work.

In 1832, when Galois was only 20 years old, he was challenged to a duel. What argument led to the challenge is not clear: some say the issue was political, while others maintain the duel was fought over a fickle lady's wavering love. The truth may never be known, but the turbulent, brilliant, and idealistic Galois died of his wounds. Fortunately for mathematics, the night before the duel he wrote down his main mathematical results and entrusted them to a friend. This time, they weren't lost—but they were only published 15 years after his death. The mathematical world was not ready for them before then!

Algebra today is organized axiomatically, and as such it is abstract. Mathematicians study algebraic structures from a general point of view, compare different structures, and find relationships between them. This abstraction and generalization might appear to be hopelessly impractical—but it is not! The general approach in algebra has produced powerful new methods for “algebraizing” different parts of mathematics and science, formulating problems which could never have been formulated before, and finding entirely new kinds of solutions.

Such excursions into pure mathematical fancy have an odd way of running ahead of physical science, providing a theoretical framework to account for facts even before those facts are fully known. This pattern is so characteristic that many mathematicians see themselves as pioneers in a world of *possibilities* rather than facts. Mathematicians study *structure* independently of content, and their science is a voyage of exploration through all the kinds of structure and order which the human mind is capable of discerning.

OPERATIONS

Addition, subtraction, multiplication, division—these and many others are familiar examples of operations on appropriate sets of numbers.

Intuitively, an operation on a set A is a way of combining any two elements of A to produce another element in the same set A .

Every operation is denoted by a symbol, such as $+$, $\times$, or $\div$. In this book we will look at operations from a lofty perspective; we will discover facts pertaining to operations *generally* rather than to specific operations on specific sets. Accordingly, we will sometimes make up operation symbols such as $*$ and $\odot$ to refer to arbitrary operations on arbitrary sets.

Let us now define formally what we mean by an *operation* on set A . Let A be *any set*:

An operation $$ on A is a rule which assigns to each ordered pair (a, b) of elements of A exactly one element $a * b$ in A .*

There are three aspects of this definition which need to be stressed:

1. *$a * b$ is defined for every ordered pair (a, b) of elements of A .* There are many rules which look deceptively like operations but are not, because this condition fails. Often $a * b$ is defined for all the obvious choices of a and b , but remains undefined in a few exceptional cases. For example, division does not qualify as an operation on the set $\mathbb{R}$ of the real numbers, for there are ordered pairs such as $(3, 0)$ whose quotient $3/0$ is undefined. In order to be an operation on $\mathbb{R}$, division would have to associate a real number ab with *every* ordered pair (a, b) of elements of $\mathbb{R}$. No exceptions allowed!
2. *$a * b$ must be uniquely defined.* In other words, the value of $a * b$ must be given unambiguously. For example, one might attempt to define an operation $\square$ on the set $\mathbb{R}$ of the real numbers by letting $a \square b$ be the number whose square is ab . Obviously this is ambiguous because $2 \square 8$, let us say, may be either 4 or -4. Thus, $\square$ does not qualify as an operation on $\mathbb{R}$!
3. *If a and b are in A , $a * b$ must be in A .* This condition is often expressed by saying that A is *closed* under the operation $*$. If we propose to define an operation $*$ on a set A , we must take care that $*$, when applied to elements of A , *does not take us out of A* . For example, division cannot be regarded as an operation on the set of the integers, for there are pairs of integers such as $(3, 4)$ whose quotient $3/4$ is not an integer.

On the other hand, division does qualify as an operation on the set of all the *positive real*

numbers, for the quotient of any two positive real numbers is a uniquely determined positive real number.

An operation is *any* rule which assigns to each ordered pair of elements of A a unique element in A . Therefore it is obvious that there are, in general, many possible operations on a given set A . If, for example, A is a set consisting of just two distinct elements, say a and b , each operation on A may be described by a table such as this one:

(x, y)	$x * y$
(a, a)	
(a, b)	
(b, a)	
(b, b)	

In the left column are listed the four possible ordered pairs of elements of A , and to the right of each pair (x, y) is the value of $x * y$. Here are a few of the possible operations:

(a, a)	a	(a, a)	a	(a, a)	b	(a, a)	b
(a, b)	a	(a, b)	b	(a, b)	a	(a, b)	b
(b, a)	a	(b, a)	a	(b, a)	b	(b, a)	b
(b, b)	a	(b, b)	b	(b, b)	a	(b, b)	a

Each of these tables describes a *different* operation on A . Each table has four rows, and each row may be filled with either an a or a b ; hence there are 16 possible ways of filling the table, corresponding to *16 possible operations* on the set A .

We have already seen that any operation on a set A comes with certain “options.” An operation $*$ may be *commutative*, that is, it may satisfy

$$a * b = b * a \quad (1)$$

for any two elements a and b in A . It may be *associative*, that is, it may satisfy the equation

$$(a * b) * c = a * (b * c) \quad (2)$$

for any three elements a , b , and c in A .

To understand the importance of the associative law, we must remember that an operation is a way of combining *two* elements; so if we want to combine *three* elements, we can do so in different ways. If we want to combine a , b , and c *without changing their order*, we may either combine a with the result of combining b and c , which produces $a * (b * c)$; or we may first combine a with b , and then combine the result with c , producing $(a * b) * c$. The associative law asserts that these two possible ways of combining three elements (without changing their order) produce the same result.

For example, the addition of real numbers is associative because $a + (b + c) = (a + b) + c$. However, division of real numbers is *not* associative: for instance, $3/(4/5)$ is $15/4$, whereas $(3/4)/5$ is $3/20$.

If there is an element e in A with the property that

$$e * a = a \quad \text{and} \quad a * e = a \quad \text{for every element } a \text{ in } A \quad (3)$$

then e is called an *identity* or “neutral” element with respect to the operation $*$. Roughly speaking, Equation (3) tells us that when e is combined with any element a , it does not change a . For example, in the set $\mathbb{R}$ of the real numbers, 0 is a neutral element for addition, and 1 is a neutral element for multiplication.

If a is any element of A , and x is an element of A such that

$$a * x = e \quad \text{and} \quad x * a = e \quad (4)$$

then x is called an *inverse* of a . Roughly speaking, Equation (4) tells us that when an element is combined with its inverse it produces the neutral element. For example, in the set $\mathbb{R}$ of the real numbers, $-a$ is the inverse of a with respect to addition; if $a \neq 0$, then $1/a$ is the inverse of a with respect to multiplication.

The inverse of a is often denoted by the symbol a^{-1} . (The symbol a^{-1} is usually pronounced “ a inverse.”)

EXERCISES

Throughout this book, the exercises are grouped into exercise sets, each set being identified by a letter A, B, C, etc, and headed by a descriptive title. Each exercise set contains six to ten exercises, numbered consecutively. *Generally, the exercises in each set are independent of each other and may be done separately.* However, when the exercises in a set are related, with some exercises building on preceding ones so that they must be done in sequence, this is indicated with a symbol t in the margin to the left of the heading.

The symbol # next to an exercise number indicates that a partial solution to that exercise is given in the Answers section at the end of the book.

A. Examples of Operations

Which of the following rules are operations on the indicated set? ($\mathbb{Z}$ designates the set of the integers, $\mathbb{Q}$ the rational numbers, and $\mathbb{R}$ the real numbers.) For each rule which is not an operation, explain why it is not.

Example $a * b = \frac{a+b}{ab}$, on the set $\mathbb{Z}$.

Solution This is not an operation on $\mathbb{Z}$. There are integers a and b such that $(a+b)/ab$ is not an integer. (For example,

$$\frac{2+3}{2 \cdot 3} = \frac{5}{6}$$

is not an integer.) Thus, $\mathbb{Z}$ is not closed under $*$.

- 1 $a * b = \sqrt{|ab|}$, on the set $\mathbb{Q}$.
- 2 $a * b = a \ln b$, on the set $\{x \in \mathbb{R} : x > 0\}$.
- 3 $a * b$ is a root of the equation $x^2 - a^2b^2 = 0$, on the set $\mathbb{R}$.
- 4 Subtraction, on the set $\mathbb{Z}$.
- 5 Subtraction, on the set $\{n \in \mathbb{Z} : \geq 0\}$.
- 6 $a * b = |a - b|$, on the set $\{n \in \mathbb{Z} : \geq 0\}$.

B. Properties of Operations

Each of the following is an operation $*$ on U . Indicate whether or not

- (i) it is commutative,
- (ii) it is associative,
- (iii) $\mathbb{R}$ has an identity element with respect to $*$,
- (iv) every $x \in \mathbb{R}$ has an inverse with respect to $*$.

Instructions For (i), compute $x * y$ and $y * x$, and verify whether or not they are equal. For (ii), compute $x * (y * z)$ and $(x * y) * z$, and verify whether or not they are equal. For (iii), first solve the equation $x * e = x$ for e ; if the equation cannot be solved, there is no identity element. If it *can* be solved, it is still necessary to check that $e * x = x * e = x$ for any $x \in \mathbb{R}$. If it checks, then e is an identity element. For (iv), first note that if there is no identity element, there can be no inverses. If there *is* an identity element e , first solve the equation $x * x' = e$ for x' if the equation cannot be solved, x does not have an inverse. If it *can* be solved, check to make sure that $x * x' = x' * x = x' * x = e$. If this checks, x' is the inverse of x .

Example $x * y = x + y + 1$

Associative		Commutative		Identity		Inverses	
Yes ☒	No ☐	Yes ☒	No ☐	Yes ☒	No ☐	Yes ☒	No ☐

- (i) $x * y = x + y + 1$; $y * x = y + x + 1 = x + y + 1$.
(Thus, $*$ is commutative.)
- (ii) $x * (y * z) = x * (y + z + 1) = x + (y + z + 1) + 1 = x + y + z + 2$.
 $(x * y) * z = (x + y + 1) * z = (x + y + 1) + z + 1 = x + y + z + 2$.
($*$ is associative.)
- (iii) Solve $x * e = x$ for e : $x * e = x + e + 1 = x$; therefore, $e = -1$.
Check: $x * (-1) = x + (-1) + 1 = x$; $(-1) * x = (-1) + x + 1 = x$.
Therefore, -1 is the identity element.
($*$ has an identity element.)
- (iv) Solve $x * x' = -1$ for x' : $x * x' = x + x' + 1 = -1$; therefore $x' = -x - 2$. Check: $x * (-x - 2) = x + (-x - 2) + 1 = -1$; $(-x - 2) * x = (-x - 2) + x + 1 = -1$. Therefore, $-x - 2$ is the inverse of x .
(Every element has an inverse.)

1 $x * y = x + 2y + 4$

Commutative		Associative		Identity		Inverses	
Yes ☐	No ☐	Yes ☐	No ☐	Yes ☐	No ☐	Yes ☐	No ☐

- (i) $x * y = x + 2y + 4$; $y * x =$
- (ii) $x * (y * z) = x * () =$
 $(x * y) * z = () * z =$
- (iii) Solve $x * e = x$ for e . Check.
- (iv) Solve $x * x' = e$ for x' . Check.

2 $x * y = x + 2y - xy$

Commutative		Associative		Identity		Inverses	
Yes ☐	No ☐	Yes ☐	No ☐	Yes ☐	No ☐	Yes ☐	No ☐

$$3 \quad x * y = |x + y|$$

<i>Commutative</i>	<i>Associative</i>	<i>Identity</i>	<i>Inverses</i>
Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐

$$4 \quad x * y = |x - y|$$

<i>Commutative</i>	<i>Associative</i>	<i>Identity</i>	<i>Inverses</i>
Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐

$$5 \quad x * y = xy + 1$$

<i>Commutative</i>	<i>Associative</i>	<i>Identity</i>	<i>Inverses</i>
Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐

$$6 \quad x * y = \max \{x, y\} = \text{the larger of the two numbers } x \text{ and } y$$

<i>Commutative</i>	<i>Associative</i>	<i>Identity</i>	<i>Inverses</i>
Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐

$$7 \quad x * y = \frac{xy}{x + y + 1} \text{ (on the set of positive real numbers)}$$

<i>Commutative</i>	<i>Associative</i>	<i>Identity</i>	<i>Inverses</i>
Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐	Yes ☐ No ☐

C. Operations on a Two-Element Set

Let A be the two-element set $A = \{a, b\}$.

1 Write the tables of all 16 operations on A . (Use the format explained on page 20.)

Label these operations 0_1 to 0_{16} . Then:

2 Identify which of the operations 0_1 to 0_{16} are commutative.

3 Identify which operations, among 0_1 to 0_{16} , are associative.

4 For which of the operations 0_1 to 0_{16} is there an identity element?

5 For which of the operations 0_1 to 0_{16} does every element have an inverse?

D. Automata: The Algebra of Input/Output Sequences

Digital computers and related machines process information which is received in the form of input sequences. An *input sequence* is a finite sequence of symbols from some alphabet A . For instance, if $A = \{0, 1\}$ (that is, if the alphabet consists of only the two symbols 0 and 1), then examples of input sequences are 011010 and 10101111. If $A = \{a, b, c\}$, then examples of input sequences are *babbcac* and *cccabaa*. *Output sequences* are defined in the same way as input sequences. The set of all sequences of symbols in the alphabet A is denoted by A^* .

There is an operation on A^* called *concatenation*: If $\mathbf{a}$ and $\mathbf{b}$ are in A^* , say $\mathbf{a} = a_1a_2 \dots a_n$ and $\mathbf{b} = b_1b_2 \dots b_m$, then

$$\mathbf{ab} = a_1a_2 \dots a_nb_1b_2 \dots b_m$$

In other words, the sequence **ab** consists of the two sequences **a** and **b** end to end. For example, in the alphabet $A = \{0,1\}$, if **a** = 1001 and **b** = 010, then **ab** = 1001010.

The symbol λ denotes the empty sequence.

- 1 Prove that the operation defined above is associative.
- 2 Explain why the operation is not commutative.
- 3 Prove that there is an identity element for this operation.

THE DEFINITION OF GROUPS

One of the simplest and most basic of all algebraic structures is the *group*. A group is defined to be a set with an operation (let us call it $*$) which is associative, has a neutral element, and for which each element has an inverse. More formally,

By a group we mean a set G with an operation $$ which satisfies the axioms:*

(G1) $*$ is associative.

(G2) There is an element e in G such that $a * e = a$ and $e * a = a$ for every element a in G .

(G3) For every element a in G , there is an element a^{-1} in G such that $a * a^{-1} = e$ and $a^{-1} * a = e$.

The group we have just defined may be represented by the symbol $\langle G, * \rangle$. This notation makes it explicit that the group consists of the *set* G and the *operation* $*$. (Remember that, in general, there are other possible operations on G , so it may not always be clear which is the group's operation unless we indicate it.) If there is no danger of confusion, we shall denote the group simply with the letter G .

The groups which come to mind most readily are found in our familiar number systems. Here are a few examples.

$\mathbf{z}$ is the symbol customarily used to denote the set

$$\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$$

of the integers. The set $\mathbf{z}$, with the operation of *addition*, is obviously a group. It is called the *additive group of the integers* and is represented by the symbol $\langle \mathbf{z}, + \rangle$. Mostly, we denote it simply by the symbol $\mathbf{z}$.

$\mathbf{Q}$ designates the set of the rational numbers (that is, quotients m/n of integers, where $n \neq 0$). This set, with the operation of addition, is called the *additive group of the rational numbers*, $\langle \mathbf{Q}, + \rangle$. Most often we denote it simply by $\mathbf{Q}$.

The symbol $\mathbf{R}$ represents the set of the real numbers. $\mathbf{R}$, with the operation of addition, is called the *additive group of the real numbers*, and is represented by $\langle \mathbf{R}, + \rangle$, or simply $\mathbf{R}$.

The set of all the *nonzero rational numbers* is represented by $\mathbf{Q}^*$. This set, with the operation of *multiplication*, is the group $\langle \mathbf{Q}^*, \cdot \rangle$, or simply $\mathbf{Q}^*$. Similarly, the set of all the *nonzero real numbers* is represented by $\mathbf{R}^*$. The set $\mathbf{R}^*$ with the operation of multiplication, is the group $\langle \mathbf{R}^*, \cdot \rangle$, or simply $\mathbf{R}^*$.

Finally, $\mathbf{Q}^{\text{pos}}$ denotes the group of all the positive rational numbers, with multiplication. $\mathbf{R}^{\text{pos}}$ denotes the group of all the positive real numbers, with multiplication.

Groups occur abundantly in nature. This statement means that a great many of the algebraic structures which can be discerned in natural phenomena turn out to be groups. Typical examples, which we shall examine later, come up in connection with the structure of crystals, patterns of symmetry, and various kinds of geometric transformations. Groups are also important because they happen to be one of the fundamental building blocks out of which more complex algebraic structures are made.

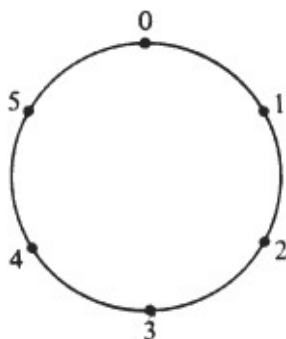
Especially important in scientific applications are the *finite* groups, that is, groups with a finite number of elements. It is not surprising that such groups occur often in applications, for in most situations of the real world we deal with only a finite number of objects.

The easiest finite groups to study are those called the *groups of integers modulo n* (where n is any positive integer greater than 1). These groups will be described in a casual way here, and a rigorous treatment deferred until later.

Let us begin with a specific example, say, the group of integers modulo 6. This group consists of a set of six elements,

$$\{0, 1, 2, 3, 4, 5\}$$

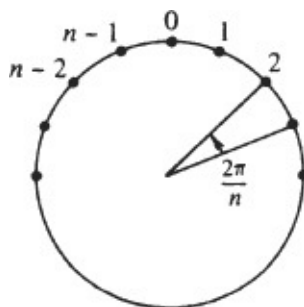
and an operation called *addition modulo 6*, which may be described as follows: Imagine the numbers 0 through 5 as being evenly distributed on the circumference of a circle. To add two numbers h and k , start with h and move clockwise k additional units around the circle: $h + k$ is where you end up. For example, $3 + 3 = 0$, $3 + 5 = 2$, and so on. The set $\{0, 1, 2, 3, 4, 5\}$ with this operation is called the *group of integers modulo 6*, and is represented by the symbol $\mathbf{z}_6$.



In general, the group of integers modulo n consists of the set

$$\{0, 1, 2, \dots, n - 1\}$$

with the operation of *addition modulo n* , which can be described exactly as previously. Imagine the numbers 0 through $n - 1$ to be points on the unit circle, each one separated from the next by an arc of length $2\pi/n$.



To add h and k , start with h and go clockwise through an arc of k times $2\pi/n$. The sum $h + k$ will, of

course, be one of the numbers 0 through $n - 1$. From geometrical considerations it is clear that this kind of addition (by successive rotations on the unit circle) is *associative*. Zero is the neutral element of this group, and $n - h$ is obviously the inverse of h [for $h + (n - h) = n$, which coincides with 0]. This group, the *group of integers modulo n* , is represented by the symbol $\mathbb{Z}_n$.

Often when working with finite groups, it is useful to draw up an “operation table.” For example, the operation table of $\mathbb{Z}_6$ is

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

The basic format of this table is as follows:

+	0	1	2	3	4	5
0						
1						
2						
3						
4						
5						

with one *row* for each element of the group and one *column* for each element of the group. Then $3 + 4$, for example, is located in the row of 3 and the column of 4. In general, any finite group $\langle G, * \rangle$ has a table

*		y	
⋮			
x		$x * y$	
⋮			

The entry in the row of x and the column of y is $x * y$.

Let us remember that the commutative law is *not* one of the axioms of group theory; hence the identity $a * b = b * a$ is not true in every group. If the commutative law holds in a group G , such a group is called a *commutative group* or, more commonly, an *abelian group*. Abelian groups are named after the mathematician Niels Abel, who was mentioned in Chapter 1 and who was a pioneer in the study of groups. All the examples of groups mentioned up to now are abelian groups, but here is an example which is not.

Let G be the group which consists of the six matrices

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix} \quad \mathbf{K} = \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix}$$

with the operation of *matrix multiplication* which was explained on page 8. This group has the following operation table, which should be checked:

	I	A	B	C	D	K
I	I	A	B	C	D	K
A	A	I	C	B	K	D
B	B	K	D	A	I	C
C	C	D	K	I	A	B
D	D	C	I	K	B	A
K	K	B	A	D	C	I

In linear algebra it is shown that the multiplication of matrices is associative. (The details are simple.) It is clear that $\mathbf{I}$ is the identity element of this group, and by looking at the table one can see that each of the six matrices in $\{\mathbf{I}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{K}\}$ has an inverse in $\{\mathbf{I}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{K}\}$. (For example, $\mathbf{B}$ is the inverse of $\mathbf{D}$, $\mathbf{A}$ is the inverse of $\mathbf{A}$, and so on.) Thus, G is a group! Now we observe that $\mathbf{AB} = \mathbf{C}$ and $\mathbf{BA} = \mathbf{K}$, so G is not commutative.

EXERCISES

A. Examples of Abelian Groups

Prove that each of the following sets, with the indicated operation, is an abelian group.

Instructions Proceed as in [Chapter 2, Exercise B](#).

1 $x * y = x + y + k$ (k a fixed constant), on the set $\mathbb{R}$ of the real numbers.

2 $x * y = \frac{xy}{2}$, on the set $\{x \in \mathbb{R} : x \neq 0\}$.

3 $x * y = x + y + xy$, on the set $\{x \in \mathbb{R} : x \neq -1\}$.

4 $x * y = \frac{x + y}{xy + 1}$, the set $\{x \in \mathbb{R} : -1 < x < 1\}$.

B. Groups on the Set $\mathbb{R} \times \mathbb{R}$

The symbol $\mathbb{R} \times \mathbb{R}$ represents the set of all ordered pairs (x, y) of real numbers. $\mathbb{R} \times \mathbb{R}$ may therefore be identified with the set of all the points in the plane. Which of the following subsets of $\mathbb{R} \times \mathbb{R}$, with the indicated operation, is a group? Which is an abelian group?

Instructions Proceed as in the preceding exercise. To find the identity element, which in these problems is an ordered pair (e_1, e_2) of real numbers, solve the equation $(a, b) * (e_1, e_2) = (a, b)$ for e_1 and e_2 . To

find the inverse (a', b') of (a, b) , solve the equation $(a, b) * (a', b') = (e_1, e_2)$ for a' and b' . [Remember that $(x, y) = (x', y')$ if and only if $x = x'$ and $y = y'$.]

1 $(a, b) * (c, d) = (ad + bc, bd)$, on the set $\{(x, y) \in \mathbb{R} \times \mathbb{R} : y \neq 0\}$.

2 $(a, b) * (c, d) = (ac, bc + d)$, on the set $\{(x, y) \in \mathbb{R} \times \mathbb{R} : x \neq 0\}$.

3 Same operation as in part 2, but on the set $\mathbb{R} \times \mathbb{R}$.

4 $(a, b) * (c, d) = (ac - bd, ad + bc)$, on the set $\mathbb{R} \times \mathbb{R}$ with the origin deleted.

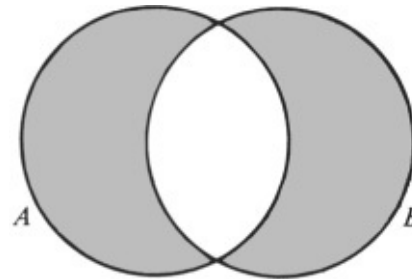
5 Consider the operation of the preceding problem on the set $\mathbb{R} \times \mathbb{R}$. Is this a group? Explain.

C. Groups of Subsets of a Set

If A and B are any two sets, their *symmetric difference* is the set $A + B$ defined as follows:

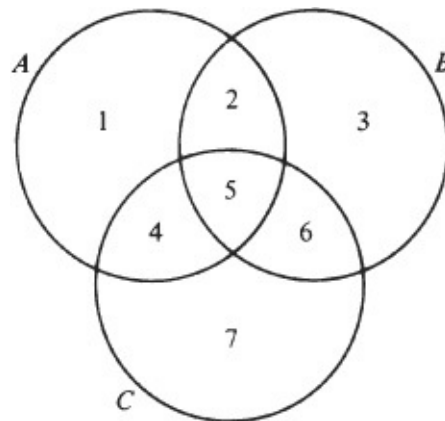
$$A + B = (A - B) \cup (B - A)$$

NOTE: $A - B$ represents the set obtained by removing from A all the elements which are in B .



The shaded area is $A + B$

It is perfectly clear that $A + B = B + A$; hence this operation is commutative. It is also associative, as the accompanying pictorial representation suggests: Let the union of A , B , and C be divided into seven regions as illustrated.



$A + B$ consists of the regions 1, 4, 3, and 6.

$B + C$ consists of the regions 2, 3, 4, and 7.

$A + (B + C)$ consists of the regions 1, 3, 5, and 7.

$(A + B) + C$ consists of the regions 1, 3, 5, and 7.

Thus, $A + (B + C) = (A + B) + C$.

If D is a set, then the *power set* of D is the set P_D of all the subsets of D . That is,

$$P_D = \{A: A \subseteq D\}$$

The operation $+$ is to be regarded as an operation on P_D .

- 1 Prove that there is an identity element with respect to the operation $+$, which is _____.
- 2 Prove every subset A of D has an inverse with respect to $+$, which is _____. Thus, $\langle P_D, + \rangle$ is a group!
- 3 Let D be the three-element set $D = \{a, b, c\}$. List the elements of P_D . (For example, one element is $\{a\}$, another is $\{a, b\}$, and so on. Do not forget the empty set and the whole set D .) Then write the operation table for $\langle P_D, + \rangle$.

D. A Checkerboard Game

1	2
3	4

Our checkerboard has only four squares, numbered 1, 2, 3, and 4. There is a single checker on the board, and it has four possible moves:

- V : Move vertically; that is, move from 1 to 3, or from 3 to 1, or from 2 to 4, or from 4 to 2.
- H : Move horizontally; that is, move from 1 to 2 or vice versa, or from 3 to 4 or vice versa.
- D : Move diagonally; that is, move from 2 to 3 or vice versa, or move from 1 to 4 or vice versa.
- I : Stay put.

We may consider an operation on the set of these four moves, which consists of performing moves successively. For example, if we move horizontally and then vertically, we end up with the same result as if we had moved diagonally:

$$H * V = D$$

If we perform two horizontal moves in succession, we end up where we started: $H * H = I$. And so on. If $G = \{V, H, D, I\}$, and $*$ is the operation we have just described, write the table of G .

*	I	V	H	D
I				
V				
H				
D				

Granting associativity, explain why $\langle G, * \rangle$ is a group.

E. A Coin Game



Imagine two coins on a table, at positions A and B . In this game there are eight possible moves:

- M_1 : Flip over the coin at A .
- M_2 : Flip over the coin at B .
- M_3 : Flip over both coins.
- M_4 : Switch the coins.
- M_5 : Flip coin at A ; then switch.
- M_6 : Flip coin at B ; then switch.
- M_7 : Flip both coins; then switch.
- I : Do not change anything.

We may consider an operation on the set $\{I, M_1, \dots, M_7\}$, which consists of performing any two moves in succession. For example, if we switch coins, then flip over the coin at A , this is the same as first flipping over the coin at B then switching:

$$M_4 * M_1 = M_2 * M_4 = M_6$$

If $G = \{I, M_1, \dots, M_7\}$ and $*$ is the operation we have just described, write the table of $\langle G, * \rangle$.

$*$	I	M_1	M_2	M_3	M_4	M_5	M_6	M_7
I								
M_1								
M_2								
M_3								
M_4								
M_5								
M_6								
M_7								

Granting associativity, explain why $\langle G, * \rangle$ is a group. Is it commutative? If not, show why not.

F. Groups in Binary Codes

The most basic way of transmitting information is to code it into strings of 0s and 1s, such as 0010111, 1010011, etc. Such strings are called *binary words*, and the number of 0s and 1s in any binary word is called its *length*. All information may be coded in this fashion.

When information is transmitted, it is sometimes received incorrectly. One of the most important purposes of coding theory is to find ways of *detecting errors*, and *correcting* errors of transmission.

If a word $\mathbf{a} = a_1a_2 \dots a_n$ is sent, but a word $\mathbf{b} = b_1b_2 \dots b_n$ is received (where the a_i and the b_j are

0s or 1s), then the *error pattern* is the word $\mathbf{e} = e_1e_2 \dots e_n$ where

$$e_i = \begin{cases} 0 & \text{if } a_i = b_i \\ 1 & \text{if } a_i \neq b_i \end{cases}$$

With this motivation, we define an operation of *adding* words, as follows: If $\mathbf{a}$ and $\mathbf{b}$ are both of length 1, we add them according to the rules

$$0 + 0 = 0 \quad 1 + 1 = 0 \quad 0 + 1 = 1 \quad 1 + 0 = 1$$

If $\mathbf{a}$ and $\mathbf{b}$ are both of length n , we add them by *adding corresponding digits*. That is (let us introduce commas for convenience),

$$(a_1, a_2, \dots, a_n) + (b_1, b_2, \dots, b_n) = (a_1 + b_1, a_2 + b_2, \dots, a_n + b_n)$$

Thus, the sum of $\mathbf{a}$ and $\mathbf{b}$ is the error pattern $\mathbf{e}$.

For example,

$$\begin{array}{r} 0010110 \\ +0011010 \\ \hline =0001100 \end{array} \qquad \begin{array}{r} 10100111 \\ +11110111 \\ \hline =01010000 \end{array}$$

The symbol $\mathbb{B}^n$ will designate the set of all the binary words of length n . We will prove that the operation of word addition has the following properties on $\mathbb{B}^n$:

1. It is commutative.
2. It is associative.
3. There is an identity element for word addition.
4. Every word has an inverse under word addition.

First, we verify the commutative law for words of length 1:

$$0 + 1 = 1 = 1 + 0$$

1 Show that $(a_1, a_2, \dots, a_n) + (b_1, b_2, \dots, b_n) = (b_1, b_2, \dots, b_n) + (a_1, a_2, \dots, a_n)$.

2 To verify the associative law, we first verify it for words of length 1:

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

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

Check the remaining six cases.

3 Show that $(a_1, \dots, a_n) + [(b_1, \dots, b_n) + (c_1, \dots, c_n)] = [(a_1, \dots, a_n) + (b_1, \dots, b_n)] + (c_1, \dots, c_n)$.

4 The identity element of $\mathbb{B}^n$, that is, the identity element for adding words of length n , is _____.

5 The inverse, with respect to word addition, of any word $(a_1, \dots, a_n)$ is _____.

6 Show that $\mathbf{a} + \mathbf{b} = \mathbf{a} - \mathbf{b}$ [where $\mathbf{a} - \mathbf{b} = \mathbf{a} + (-\mathbf{b}^*)$].

7 If $\mathbf{a} + \mathbf{b} = \mathbf{c}$, show that $\mathbf{a} = \mathbf{b} + \mathbf{c}$.

G. Theory of Coding: Maximum-Likelihood Decoding

We continue the discussion started in [Exercise F](#): Recall that $\mathbb{B}^n$ designates the set of all binary words of length n . By a *code* we mean a subset of $\mathbb{B}^n$. For example, below is a code in $\mathbb{B}^5$. The code, which we shall call C_1 , consists of the following binary words of length 5:

00000
00111
01001
01110
10011
10100
11010
11101

Note that there are 32 possible words of length 5, but only eight of them are in the code C_1 . These eight words are called *codewords*; the remaining words of $\mathbb{B}^5$ are not codewords. Only codewords are transmitted. If a word is received which is not a codeword, it is clear that there has been an *error of transmission*. In a well-designed code, it is unlikely that an error in transmitting a codeword will produce another codeword (if that were to happen, the error would not be detected). Moreover, in a good code it should be fairly easy to locate errors and correct them. These ideas are made precise in the discussion which follows.

The *weight* of a binary word is the number of 1s in the word: for example, 11011 has weight 4. The *distance* between two binary words is the number of positions in which the words differ. Thus, the distance between 11011 and 01001 is 2 (since these words differ only in their first and fourth positions). The *minimum distance* in a code is the smallest distance among all the distances between pairs of codewords. For the code C_1 , above, pairwise comparison of the words shows that the minimum distance is 2. What this means is that *at least two* errors of transmission are needed in order to transform a codeword into another codeword; single errors will change a codeword into a *noncodeword*, and the error will therefore be detected. In more desirable codes (for example, the so-called Hamming code), the minimum distance is 3, so any one or two errors are *always* detected, and only three errors in a single word (a very unlikely occurrence) might go undetected.

In practice, a code is constructed as follows: in every codeword, certain positions are *information positions*, and the remaining positions are *redundancy positions*. For instance, in our code C_1 , the first three positions of every codeword are the information positions: if you look at the eight codewords (and confine your attention only to the first three digits in each word), you will see that every three-digit sequence of 0s and 1s is there namely,

000, 001, 010, 011, 100, 101, 110, 111

The numbers in the fourth and fifth positions of every codeword satisfy *parity-check equations*.

1 Verify that every codeword $a_1a_2a_3a_4a_5$ in C_1 satisfies the following two parity-check equations: $a_4 = a_1 + a_3$; $a_5 = a_1 + a_2 + a_3$.

2 Let C_2 be the following code in $\mathbb{B}^6$. The first three positions are the information positions, and every codeword $a_1a_2a_3a_4a_5a_6$ satisfies the parity-check equations $a_4 = a_2$, $a_5 = a_1 + a_2$, and $a_6 = a_1 + a_2 + a_3$.

(a) List the codewords of C_2 .

(b) Find the minimum distance of the code C_2 .

(c) How many errors in any codeword of C_2 are sure to be detected? Explain.

3 Design a code in $\mathbb{B}^4$ where the first two positions are information positions. Give the parity-check equations, list the codewords, and find the minimum distance.

If $\mathbf{a}$ and $\mathbf{b}$ are any two words, let $d(\mathbf{a}, \mathbf{b})$ denote the distance between $\mathbf{a}$ and $\mathbf{b}$. To *decode* a received word $\mathbf{x}$ (which may contain errors of transmission) means to find the codeword closest to $\mathbf{x}$, that is, the codeword $\mathbf{a}$ such that $d(\mathbf{a}, \mathbf{x})$ is a minimum. This is called *maximum-likelihood decoding*.

4 Decode the following words in C_1 : 11111, 00101, 11000, 10011, 10001, and 10111.

You may have noticed that the last two words in part 4 had ambiguous decodings: for example, 10111 may be decoded as either 10011 or 00111. This situation is clearly unsatisfactory. We shall see next what conditions will ensure that every word can be decoded into only *one* possible codeword.

In the remaining exercises, let C be a code in $\mathbb{B}^n$, let m denote the minimum distance in C , and let $\mathbf{a}$ and $\mathbf{b}$ denote codewords in C .

5 Prove that it is possible to detect up to $m - 1$ errors. (That is, if there are errors of transmission in $m - 1$ or fewer positions of a codeword, it can always be determined that the received word is incorrect.)

6 By the *sphere of radius k* about a codeword $\mathbf{a}$ we mean the set of all words in $\mathbb{B}^n$ whose distance from $\mathbf{a}$ is no greater than k . This set is denoted by $S_k(\mathbf{a})$; hence

$$S_k(\mathbf{a}) = \{\mathbf{x}: d(\mathbf{a}, \mathbf{x}) \leq k\}$$

If $t = \frac{1}{2}(m - 1)$, prove that any two spheres of radius t , say $S_t(\mathbf{a})$ and $S_t(\mathbf{b})$, have no elements in common.

[HINT: Assume there is a word $\mathbf{x}$ such that $\mathbf{x} \in S_t(\mathbf{a})$ and $\mathbf{x} \in S_t(\mathbf{b})$. Using the definitions of t and m , show that this is impossible.]

7 Deduce from part 6 that if there are t or fewer errors of transmission in a codeword, the received word will be decoded correctly.

8 Let C_2 be the code described in part 2. (If you have not yet found the minimum distance in C_2 , do so now.) Using the results of parts 5 and 7, explain why two errors in any codeword can always be detected, and why one error in any codeword can always be corrected.

CHAPTER FOUR

ELEMENTARY PROPERTIES OF GROUPS

Is it possible for a group to have *two different* identity elements? Well, suppose e_1 and e_2 are identity elements of some group G . Then

$$e_1 * e_2 = e_2 \quad \text{because } e_1 \text{ is an identity element, and}$$

$$e_1 * e_2 = e_1 \quad \text{because } e_2 \text{ is an identity element}$$

Therefore

$$e_1 = e_2$$

This shows that in every group there is *exactly one* identity element.

Can an element a in a group have *two different inverses*? Well, if a_1 and a_2 are both inverses of a , then

$$a_1 * (a * a_2) = a_1 * e = a_1$$

and

$$(a_1 * a) * a_2 = e * a_2 = a_2$$

By the associative law, $a_1 * (a * a_2) = (a_1 * a) * a_2$; hence $a_1 = a_2$. This shows that in every group, each element has *exactly one* inverse.

Up to now we have used the symbol $*$ to designate the group operation. Other, more commonly used symbols are $+$ and $\cdot$ (“plus” and “multiply”). When $+$ is used to denote the group operation, we say we are using *additive notation*, and we refer to $a + b$ as the *sum* of a and b . (Remember that a and b do not have to be numbers and therefore “sum” does not, in general, refer to adding numbers.) When $\cdot$ is used to denote the group operation, we say we are using *multiplicative notation*, we usually write ab instead of $a \cdot b$, and call ab the *product* of a and b . (Once again, remember that “product” does not, in

general, refer to multiplying numbers.) Multiplicative notation is the most popular because it is simple and saves space. In the remainder of this book multiplicative notation will be used except where otherwise indicated. In particular, when we represent a group by a letter such as G or H , it will be understood that the group's operation is written as multiplication.

There is common agreement that in additive notation the identity element is denoted by 0, and the inverse of a is written as $-a$. (It is called the *negative* of a .) In multiplicative notation the identity element is e and the inverse of a is written as a^{-1} (" a inverse"). It is also a tradition that $+$ is to be used only for commutative operations.

The most basic rule of calculation in groups is the *cancellation law*, which allows us to cancel the factor a in the equations $ab = ac$ and $ab = ca$. This will be our first theorem about groups.

Theorem 1 *If G is a group and a, b, c are elements of G , then*

- (i) $ab = ac$ implies $b = c$ and
- (ii) $ba = ca$ implies $b = c$

It is easy to see why this is true: if we multiply (on the left) both sides of the equation $ab = ac$ by a^{-1} , we get $b = c$. In the case of $ba = ca$, we multiply on the right by a^{-1} . This is the *idea* of the proof; now here is the proof:

Suppose	$ab = ac$
Then	$a^{-1}(ab) = a^{-1}(ac)$
By the associative law,	$(a^{-1}a)b = (a^{-1}a)c$
that is,	$eb = ec$
Thus, finally,	$b = c$

Part (ii) is proved analogously.

In general, we *cannot* cancel a in the equation $ab = ca$. (Why not?)

Theorem 2 *If G is a group and a, b are elements of G , then*

$$ab=e \text{ implies } a=b^{-1} \text{ and } b=a^{-1}$$

The proof is very simple: if $ab = e$, then $ab = aa^{-1}$ so by the cancellation law, $b = a^{-1}$. Analogously, $a = b^{-1}$.

This theorem tells us that if the product of two elements is equal to e , these elements are inverses of each other. In particular, if a is the inverse of b , then b is the inverse of a .

The next theorem gives us important information about computing inverses.

Theorem 3 *If G is a group and a, b are elements of G , then*

- (i) $(ab^{-1} = b^{-1}a^{-1}$ and
- (ii) $(a^{-1})^{-1}=a$

The first formula tells us that the inverse of a product is the product of the inverses in reverse order. The next formula tells us that a is the inverse of the inverse of a . The proof of (i) is as follows:

$$\begin{aligned}
(ab)(b^{-1}a^{-1}) &= a[(bb^{-1})a^{-1}] && \text{by the associative law} \\
&= a[ea^{-1}] && \text{because } bb^{-1} = e \\
&= aa^{-1} \\
&= e
\end{aligned}$$

Since the product of ab and $b^{-1}a^{-1}$ is equal to e , it follows by [Theorem 2](#) that they are each other's inverses. Thus, $(ab)^{-1} = b^{-1}a^{-1}$. The proof of (ii) is analogous but simpler: $aa^{-1} = e$, so by [Theorem 2](#) a is the inverse of a^{-1} , that is, $a = (a^{-1})^{-1}$.

The associative law states that the two products $a(bc)$ and $(ab)c$ are equal; for this reason, no confusion can result if we denote either of these products by writing abc (without any parentheses), and call abc the product of these *three* elements in this order.

We may next define the product of any *four* elements a, b, c , and d in G by

$$abcd = a(bcd)$$

By successive uses of the associative law we find that

$$a(bc)d = ab(cd) = (ab)(cd) = (ab)cd$$

Hence the product $abcd$ (without parentheses, but without changing the order of its factors) is defined without ambiguity.

In general, any two products, each involving the same factors in the same order, are equal. The net effect of the associative law is that *parentheses are redundant*.

Having made this observation, we may feel free to use products of several factors, such as $a_1a_2 \cdots a_n$, without parentheses, whenever it is convenient. Incidentally, by using the identity $(ab)^{-1} = b^{-1}a^{-1}$ repeatedly, we find that

$$(a_1a_2 \cdots a_n)^{-1} = a_n^{-1} \cdots a_2^{-1}a_1^{-1}$$

If G is a finite group, the number of elements in G is called the *order* of G . It is customary to denote the order of G by the symbol

$$|G|$$

EXERCISES

Remark on notation In the exercises below, the exponential notation a^n is used in the following sense: if a is any element of a group G , then a^2 means aa , a^3 means aaa , and, in general, a^n is the product of n factors of a , for any positive integer n .

A. Solving Equations in Groups

Let a, b, c , and x be elements of a group G . In each of the following, solve for x in terms of a, b , and c .

Example Solve simultaneously: $x^2 = b$ and $x^5 = e$

From the first equation, $b = x^2$

Squaring, $b^2 = x^4$

Multiplying on the left by x , $xb^2 = xx^4 = x^5 = e$. (Note: $x^5 = e$ was given.)

Multiplying by $(b^2)^{-1}$, $xb^2(b^2)^{-1}$. Therefore, $x = (b^2)^{-1}$.

Solve:

1 $axb = c$

2 $x^2b = xa^{-1}c$

Solve simultaneously:

3 $x^2a = bxc^{-1}$ and $acx = xac$

4 $ax^2 = b$ and $x^3 = e$

5 $x^2 = a^2$ and $x^5 = e$

6 $(xax)^3 = bx$ and $x^2a = (xa)^{-1}$

B. Rules of Algebra in Groups

For each of the following rules, either prove that it is true in every group G , or give a counterexample to show that it is false in some groups. (All the counterexamples you need may be found in the group of matrices $\{I, A, B, C, D, K\}$ described on page 28.)

1 If $x^2 = e$, then $x = e$.

2 If $x^2 = a^2$, then $x = a$.

3 $(ab)^2 = a^2b^2$

4 If $x^2 = x$, then $x = e$.

5 For every $x \in G$, there is some $y \in G$ such that $x = y^2$. (This is the same as saying that every element of G has a “square root.”)

6 For any two elements x and y in G , there is an element z in G such that $y = xz$.

C. Elements That Commute

If a and b are in G and $ab = ba$, we say that a and b *commute*. Assuming that a and b commute, prove the following:

1 a^{-1} and b^{-1} commute.

2 a and b^{-1} commute. (HINT: First show that $a = b^{-1}ab$.)

3 a commutes with ab .

4 a^2 commutes with b^2 .

5 xax^{-1} commutes with xbx^{-1} , for any $x \in G$.

6 $ab = ba$ iff $aba^{-1} = b$.

(The abbreviation iff stands for “if and only if.” Thus, first prove that if $ab = ba$, then $aba^{-1} = b$. Next, prove that if $aba^{-1} = b$, then $ab = ba$. Proceed roughly as in [Exercise A](#). Thus, assuming $ab = ba$, solve for b . Next, assuming $aba^{-1} = b$, solve for ab .)

7 $ab = ba$ iff $aba^{-1}b^{-1} = e$.

† D. Group Elements and Their Inverses¹

Let G be a group. Let a, b, c denote elements of G , and let e be the neutral element of G .

1 Prove that if $ab = e$, then $ba = e$. (HINT: See [Theorem 2](#).)

2 Prove that if $abc = e$, then $cab = e$ and $bca = e$.

3 State a generalization of parts 1 and 2

Prove the following:

4 If $xay = a^{-1}$, then $yax = a^{-1}$.

5 Let a, b , and c each be equal to its own inverse. If $ab = c$, then $bc = a$ and $ca = b$.

6 If abc is its own inverse, then bca is its own inverse, and cab is its own inverse.

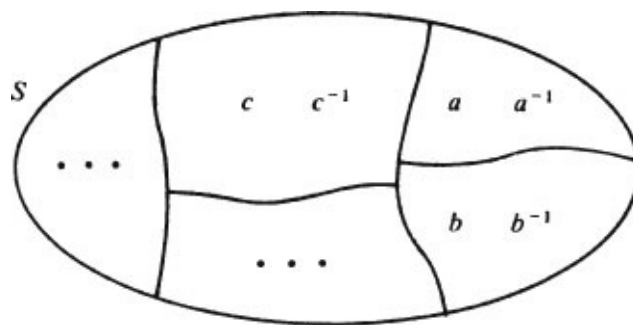
7 Let a and b each be equal to its own inverse. Then ba is the inverse of ab .

8 $a = a^{-1}$ iff $aa = e$. (That is, a is its own inverse iff $a^2 = e$.)

9 Let $c = c^{-1}$. Then $ab = c$ iff $xy^2 abc = e$.

† E. Counting Elements and Their Inverses

Let G be a finite group, and let S be the set of all the elements of G which are *not* equal to their own inverse. That is, $S = \{x \in G : x \neq x^{-1}\}$. The set S can be divided up into pairs so that each element is paired off with its own inverse. (See diagram on the next page.) Prove the following:



1 In any finite group G , the number of elements not equal to their own inverse is an even number.

2 The number of elements of G equal to their own inverse is odd or even, depending on whether the number of elements in G is odd or even.

3 If the order of G is even, there is at least one element x in G such that $x \neq e$ and $x = x^{-1}$.

In parts 4 to 6, let G be a finite *abelian* group, say, $G = \{e, a_1, a_2, \dots, a_n\}$. Prove the following:

4 $(a_1 a_2 \cdots a_n)^2 = e$

5 If there is no element $x \neq e$ in G such that $x = x^{-1}$, then $a_1 a_2 \cdots a_n = e$.

6 If there is exactly one $x \neq e$ in G such that $x = x^{-1}$, then $a_1 a_2 \cdots a_n = x$.

† F. Constructing Small Groups

In each of the following, let G be any group. Let e denote the neutral element of G .

1 If a, b are any elements of G , prove each of the following:

(a) If $a^2 = a$, then $a = e$.

(b) If $ab = a$, then $b = e$.

(c) If $ab = b$, then $a = e$.

2 Explain why every row of a group table must contain each element of the group exactly once. (HINT: Suppose jc appears twice in the row of a :

	$\cdots$	y_1	$\cdots$	y_2	$\cdots$
$\vdots$		$\vdots$		$\vdots$	
a	$\cdots$	x	$\cdots$	x	$\cdots$

Now use the cancellation law for groups.)

3 There is *exactly one group* on any set of three distinct elements, say the set $\{e, a, b\}$. Indeed, keeping in mind parts 1 and 2 above, there is only one way of completing the following table. Do so! *You need not prove associativity.*

	e	a	b
e	e	a	b
a	a		
b	b		

4 There is exactly one group G of four elements, say $G = \{e, a, b, c\}$, satisfying the additional property that $xx = e$ for every $x \in G$. Using only part 1 above, complete the following group table of G :

	e	a	b	c
e	e	a	b	c
a	a			
b	b			
c	c			

5 There is exactly one group G of four elements, say $G = \{e, a, b, c\}$, such that $xx = e$ for some $x \neq e$ in G , and $yy \neq e$ for some $y \in G$ (say, $aa = e$ and $bb \neq e$). Complete the group table of G , as in the preceding exercise.

6 Use [Exercise E3](#) to explain why the groups in parts 4 and 5 are the only possible groups of four elements (except for renaming the elements with different symbols).

G. Direct Products of Groups

If G and H are any two groups, their *direct product* is a new group, denoted by $G \times H$, and defined as follows: $G \times H$ consists of all the ordered pairs (x, y) where x is in G and y is in H . That is,

$$G \times H = \{(x, y) : x \in G \text{ and } y \in H\}$$

The operation of $G \times H$ consists of multiplying corresponding components:

$$(x, y) \cdot (x' y') = (xx', yy')$$

If G and H are denoted additively, it is customary to denote $G \times H$ additively:

$$(x, y) + (x' y') = (x+x', y+y')$$

1 Prove that $G \times H$ is a group by checking the three group axioms, (G1) to (G3):

$$(G1) \quad (x_1, y_1)[(x_2, y_2)(x_3, y_3)] = (\quad , \quad)$$

$$[(x_1, y_1)(x_2, y_2)](x_3, y_3) = (\quad , \quad)$$

$$(G2) \quad \text{Let } e_G \text{ be the identity element of } G, \text{ and } e_H \text{ the identity element of } H.$$

The identity element of $G \times H$ is $(\quad, \quad)$. Check

$$(G3) \text{ For each } (a, b) \in G \times H, \text{ the inverse of } (a, b) \text{ is } (\quad, \quad). \text{ Check.}$$

2 List the elements of $\mathbf{Z}_2 \times \mathbf{Z}_3$, and write its operation table. (NOTE: There are six elements, each of which is an ordered pair. The notation is additive.)

3 If G and H are abelian, prove that $G \times H$ is abelian.

4 Suppose the groups G and H both have the following property:

Every element of the group is its own inverse.

Prove that $G \times H$ also has this property.

H. Powers and Roots of Group Elements

Let G be a group, and $a, b \in G$. For any positive integer n we define a^n by

$$a^n = \underbrace{aaa \cdots a}_{n \text{ factors}}$$

If there is an element $x \in G$ such that $a = x^2$, we say that a has a square root in G . Similarly, if $a = y^3$ for some $y \in G$, we say a has a cube root in G . In general, a has an n th root in G if $a = z^n$ for some $z \in G$. Prove the following:

1 $(bab^{-1})^n = ba^n b^{-1}$, for every positive integer n . Prove by induction. (Remember that to prove a formula such as this one by induction, you first prove it for $n = 1$; next you prove that *if* it is true for $n = k$, *then* it must be true for $n = k + 1$. You may conclude that it is true for every positive integer n . Induction is explained more fully in Appendix C.)

2 If $ab = ba$, then $(ab)^n = a^n b^n$ for every positive integer n . Prove by induction.

3 If $xax = e$, then $(xa)^{2n} = a^n$.

4 If $a^3 = e$, then a has a square root.

5 If $a^2 = e$, then a has a cube root.

6 If $a \in 1$ has a cube root, so does a .

7 If $x^2 ax = a^{-1}$, then a has a cube root. (HINT: Show that xax is a cube root of a^{-1} .)

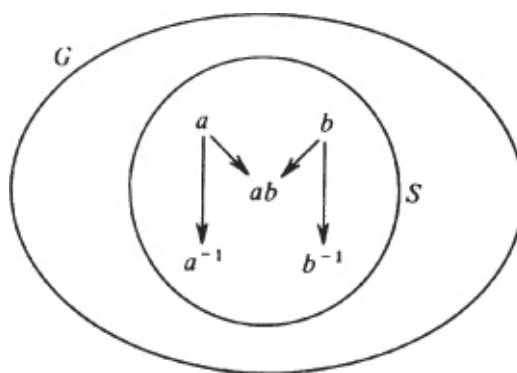
8 If $xax = b$, then 06 has a square root.

¹ When the exercises in a set are related, with some exercises building on preceding ones so that they must be done in sequence, this is indicated with a symbol t in the margin to the left of the heading.

CHAPTER FIVE

SUBGROUPS

Let G be a group, and S a nonempty subset of G . It may happen (though it doesn't have to) that the product of every pair of elements of S is in S . If it happens, we say that S is *closed with respect to multiplication*. Then, it may happen that the inverse of every element of S is in S . In that case, we say that S is *closed with respect to inverses*. If both these things happen, we call S a *subgroup* of G .



When the operation of G is denoted by the symbol $+$, the wording of these definitions must be adjusted: if the *sum* of every pair of elements of S is in S , we say that S is *closed with respect to addition*. If the negative of every element of S is in S , we say that S is *closed with respect to negatives*. If both these things happen, S is a *subgroup* of G .

For example, the *set of all the even integers* is a subgroup of the additive group $\mathbf{Z}$ of the integers. Indeed, the sum of any two even integers is an even integer, and the negative of any even integer is an even integer.

As another example, $\mathbf{Q}^*$ (the group of the nonzero rational numbers, under multiplication) is a subgroup of $\mathbf{R}^*$ (the group of the nonzero real numbers, under multiplication). Indeed, $\mathbf{Q}^* \subseteq \mathbf{R}^*$ because every rational number is a real number. Furthermore, the product of any two rational numbers is rational, and the inverse (that is, the reciprocal) of any rational number is a rational number.

An important point to be noted is this: if S is a subgroup of G , *the operation of S is the same as the operation of G* . In other words, if a and b are elements of S , the product ab computed in S is precisely the product ab computed in G .

For example, it would be meaningless to say that $\langle \mathbb{Q}^*, \cdot \rangle$ is a subgroup of $\langle \mathbb{R}, + \rangle$; for although it is true that $\mathbb{Q}^*$ is a subset of $\mathbb{R}$, the operations on these two groups are different.

The importance of the notion of subgroup stems from the following fact: *if G is a group and S is a subgroup of G , then S itself is a group.*

It is easy to see why this is true. To begin with, the operation of G , restricted to elements of S , is certainly an operation on S . *It is associative:* for if a , b , and c are in S , they are in G (because $S \subseteq G$); but G is a group, so $a(bc) = (ab)c$. Next, *the identity element e of G is in S* (and continues to be an identity element in S) for S is nonempty, so S contains an element a ; but S is closed with respect to inverses, so S also contains a^{-1} ; thus, S contains $aa^{-1} = e$, because S is closed with respect to multiplication. Finally, *every element of S has an inverse in S* because S is closed with respect to inverses. Thus, S is a group!

One reason why the notion of subgroup is useful is that it provides us with an easy way of showing that certain things are groups. Indeed, if G is already known to be a group, and S is a subgroup of G , we may conclude that S is a group without having to check all the items in the definition of “group.” This conclusion is illustrated by the next example.

Many of the groups we use in mathematics are groups whose elements are functions. In fact, historically, the first groups ever studied as such were groups of functions.

$\mathcal{F}(\mathbb{R})$ represents the *set of all functions from $\mathbb{R}$ to $\mathbb{R}$* , that is, the set of all real-valued functions of a real variable. In calculus we learned how to add functions: if f and g are functions from $\mathbb{R}$ to $\mathbb{R}$, their *sum* is the function $f + g$ given by

$$[f + g](x) = f(x) + g(x) \quad \text{for every real number } x$$

Clearly, $f + g$ is again a function from $\mathbb{R}$ to $\mathbb{R}$, and is uniquely determined by f and g .

$\mathcal{F}(\mathbb{R})$, with the operation $+$ for adding functions, is the group $\langle \mathcal{F}(\mathbb{R}), + \rangle$, or simply $\mathcal{F}(\mathbb{R})$. The details are simple, but first, let us remember what it means for two functions to be equal. If f and g are functions from $\mathbb{R}$ to $\mathbb{R}$, then f and g are *equal* (that is, $f = g$) if and only if $f(x) = g(x)$ for every real number x . In other words, to be equal, f and g must yield the same value when applied to every real number x .

To check that $+$ is associative, we must show that $f + [g + h] = [f + g] + h$, for every three functions, f , g , and h in $\mathcal{F}(\mathbb{R})$. This means that for any real number x , $\{f + [g + h]\}(x) = \{[f + g] + h\}(x)$. Well,

$$\{f + [g + h]\}(x) = f(x) + [g + h](x) = f(x) + g(x) + h(x)$$

and $\{[f + g] + h\}(x)$ has the same value.

The neutral element of $\mathcal{F}(\mathbb{R})$ is the function $\mathbf{0}$ given by

$$\mathbf{0}(x) = 0 \quad \text{for every real number } x$$

To show that $\mathbf{0} + f = f$, one must show that $[\mathbf{0} + f](x) = f(x)$ for every real number x . This is true because $[\mathbf{0} + f](x) = \mathbf{0}(x) + f(x) = 0 + f(x) = f(x)$.

Finally, the inverse of any function f is the function $-f$ given by

$$[-f](x) = -f(x) \quad \text{for every real number } x$$

One perceives immediately that $f + [-f] = \mathbf{0}$, for every function f .

$\mathcal{C}(\mathbb{R})$ represents the set of all *continuous* functions from $\mathbb{R}$ to $\mathbb{R}$. Now, $\mathcal{C}(\mathbb{R})$, with the operation $+$, is a subgroup of $\mathcal{F}(\mathbb{R})$, because we know from calculus that the sum of any two continuous functions is a continuous function, and the negative $-f$ of any continuous function f is a continuous function. Because any subgroup of a group is itself a group, we may conclude that $\mathcal{C}(\mathbb{R})$, with the operation $+$, is a group. It is denoted by $\langle \mathcal{C}(\mathbb{R}), + \rangle$, or simply $\mathcal{C}(\mathbb{R})$.

$\mathcal{D}(\mathbb{R})$ represents the set of all the *differentiable* functions from $\mathbb{R}$ to $\mathbb{R}$. It is a subgroup of $\mathcal{F}(\mathbb{R})$ because the sum of any two differentiable functions is differentiable, and the negative of any differentiable function is differentiable. Thus, $\mathcal{D}(\mathbb{R})$, with the operation of adding functions, is a group denoted by $\langle \mathcal{D}(\mathbb{R}), + \rangle$, or simply $\mathcal{D}(\mathbb{R})$.

By the way, in any group G the one-element subset $\{e\}$, containing only the neutral element, is a subgroup. It is closed with respect to multiplication because $ee = e$, and closed with respect to inverses because $e^{-1} = e$. At the other extreme, the whole group G is obviously a subgroup of itself. These two examples are, respectively, the smallest and largest possible subgroups of G . They are called the *trivial* subgroups of G . All the other subgroups of G are called *proper* subgroups.

Suppose G is a group and a, b , and c are elements of G . Define S to be the subset of G which contains *all the possible products of a, b, c , and their inverses*, in any order, with repetition of factors permitted. Thus, typical elements of S would be

$$abac^{-1}$$

$$c^{-1}a^{-1}bbc$$

and so on. It is easy to see that S is a subgroup of G : for if two elements of S are multiplied together, they yield an element of S , and the inverse of any element of S is an element of S . For example, the product of aba and $cb^{-1}ac$ is

$$abacb^{-1}ac$$

and the inverse of $ab^{-1}c^{-1}a$ is

$$a^{-1}cba^{-1}$$

S is called the *subgroup of G generated by a, b , and c* .

If $a_1, \dots, a_n$ are any finite number of elements of G , we may define the *subgroup generated by $a_1, \dots, a_n$* in the same way. In particular, if a is a single element of G , we may consider the subgroup generated by a . This subgroup is designated by the symbol $\langle a \rangle$, and is called a *cyclic subgroup* of G ; a is called its *generator*. Note that $\langle a \rangle$ consists of all the possible products of a and a^{-1} , for example, $a^{-1}aaa^{-1}$ and $aaa^{-1}aa^{-1}$. However, since factors of a^{-1} cancel factors of a , there is no need to consider products involving both a and a^{-1} side by side. Thus, $\langle a \rangle$ contains

$$a, aa, aaa, \dots,$$

$$a^{-1}, a^{-1}a^{-1}, a^{-1}a^{-1}a^{-1}, \dots,$$

as well as $aa^{-1} = e$.

If the operation of G is denoted by $+$, the same definitions can be given with “sums” instead of “products.”

In the group of matrices whose table appears on page 28, the subgroup generated by D is $\langle D \rangle = \{I, B, D\}$ and the subgroup generated by A is $\langle A \rangle = \{I, A\}$. (The student should check the table to verify this.) In fact, the entire group G of that example is generated by the two elements A and B .

If a group G is generated by a single element a , we call G a *cyclic group*, and write $G = \langle a \rangle$. For example, the additive group $\mathbf{Z}_6$ is cyclic. (What is its generator?)

Every finite group G is generated by one or more of its elements (obviously). A set of equations, involving only the generators and their inverses, is called a set of *defining equations* for G if these equations completely determine the multiplication table of G .

For example, let G be the group $\{e, a, b, b^2, ab, ab^2\}$ whose generators a and b satisfy the equations

$$a^2 = e \quad b^3 = e \quad ba = ab^2 \quad (1)$$

These three equations do indeed determine the multiplication table of G . To see this, note first that the equation $ba = ab^2$ allows us to switch powers of a with powers of b , bringing powers of a to the left, and powers of b to the right. For example, to find the product of ab and ab^2 , we compute as follows:

$$(ab)(ab^2) = \underbrace{abab^2}_{=ab^2} = aab^2b^2 = a^2b^4$$

But by [Equations \(1\)](#), $a^2 = e$ and $b^4 = b^3b = b$; so finally, $(ab)(ab^2) = b$. All the entries in the table of G may be computed in the same fashion.

When a group is determined by a set of generators and defining equations, its structure can be efficiently represented in a diagram called a *Cayley diagram*. These diagrams are explained in [Exercise G](#).

EXERCISES

A. Recognizing Subgroups

In parts 1–6 below, determine whether or not H is a subgroup of G . (Assume that the operation of H is the same as that of G .)

Instructions If H is a subgroup of G , show that both conditions in the definition of “subgroup” are satisfied. If H is *not* a subgroup of G , explain which condition fails.

Example $G = \mathbf{R}^*$, the multiplicative group of the real numbers.

$$H = \{2^n : n \in \mathbf{Z}\} \quad H \text{ is } \boxed{\times} \quad \text{is not } \boxed{\square} \quad \text{a subgroup of } G.$$

(i) If $2^n, 2^m \in H$, then $2^n 2^m = 2^{n+m}$. But $n + m \in \mathbf{Z}$, so $2^{n+m} \in H$.

(ii) If $2^n \in H$, then $1/2^n = 2^{-n}$. But $-n \in \mathbf{Z}$, so $2^{-n} \in H$.

(Note that in this example the operation of G and H is multiplication. In the next problem, it is addition.)

1 $G = \langle \mathbb{R}, + \rangle$, $H = \{\log a : a \in \mathbb{Q}, a > 0\}$. H is ☐ is not ☐ a subgroup of G .

2 $G = \langle \mathbb{R}, + \rangle$, $H = \{\log n : n \in \mathbb{Z}, n > 0\}$. H is ☐ is not ☐ a subgroup of G .

3 $G = \langle \mathbb{R}, + \rangle$, $H = \{x \in \mathbb{R} : \tan x \in \mathbb{Q}\}$. H is ☐ is not ☐ a subgroup of G .

HINT: Use the following formula from trigonometry:

$$\tan(x + y) = \frac{\tan x + \tan y}{1 - \tan x \tan y}$$

4 $G = \langle \mathbb{R}^*, \cdot \rangle$, $H = \{2^n 3^m : m, n \in \mathbb{Z}\}$. H is ☐ is not ☐ a subgroup of G .

5 $G = \langle \mathbb{R} \times \mathbb{R}, + \rangle$, $H = \{(x, y) : y = 2x\}$. H is ☐ is not ☐ a subgroup of G .

6 $G = \langle \mathbb{R} \times \mathbb{R}, + \rangle$, $H = \{(x, y) : x^2 + y^2 > 0\}$. H is ☐ is not ☐ a subgroup of G .

7 Let C and D be sets, with $C \subseteq D$. Prove that P_C is a subgroup of P_D . (See [Chapter 3, Exercise C](#).)

B. Subgroups of Functions

In each of the following, show that H is a subgroup of G .

Example $G = \langle \mathcal{F}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{F}(\mathbb{R}) : f(0) = 0\}$

(i) Suppose $f, g \in H$; then $f(0) = 0$ and $g(0) = 0$, so $[f + g](0) = f(0) + g(0) = 0 + 0 = 0$. Thus, $f + g \in H$.

(ii) If $f \in H$, then $f(0) = 0$. Thus, $[-f](0) = -f(0) = -0 = 0$, so $-f \in H$.

1 $G = \langle \mathcal{F}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{F}(\mathbb{R}) : f(x) = 0 \text{ for every } x \in [0, 1]\}$

2 $G = \langle \mathcal{F}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{F}(\mathbb{R}) : f(-x) = -f(x)\}$

3 $G = \langle \mathcal{F}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{F}(\mathbb{R}) : f \text{ is periodic of period } \pi\}$

REMARK: A function f is said to be *periodic* of period a if there is a number a , called the period of f , such that $f(x) = f(x + na)$ for every $x \in \mathbb{R}$ and $n \in \mathbb{Z}$.

4 $G = \langle \mathcal{C}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{C}(\mathbb{R}) : \int_0^1 f(x) dx = 0\}$

5 $G = \langle \mathcal{D}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{D}(\mathbb{R}) : df/dx \text{ is constant}\}$

6 $G = \langle \mathcal{F}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{F}(\mathbb{R}) : f(x) \in \mathbb{Z} \text{ for every } x \in \mathbb{R}\}$

C. Subgroups of Abelian Groups

In the following exercises, let G be an abelian group.

1 If $H = \{x \in G : x = x^{-1}\}$, that is, H consists of all the elements of G which are their own inverses, prove that H is a subgroup of G .

2 Let n be a fixed integer, and let $H = \{x \in G : x^n = e\}$. Prove that H is a subgroup of G .

3 Let $H = \{x \in G : x = y^2 \text{ for some } y \in G\}$; that is, let H be the set of all the elements of G which have a square root. Prove that H is a subgroup of G .

4 Let H be a subgroup of G , and let $K = \{x \in G : x^2 \in H\}$. Prove that K is a subgroup of G .

5. Let H be a subgroup of G , and let K consist of all the elements x in G such that some power of x is in H . That is, $K = \{x \in G : \text{for some integer } n > 0, x^n \in H\}$. Prove that K is a subgroup of G .

6 Suppose H and K are subgroups of G , and define HK as follows:

$$HK = \{xy : x \in H \text{ and } y \in K\}$$

Prove that HK is a subgroup of G .

7 Explain why parts 4–6 are not true if G is not abelian.

D. Subgroups of an Arbitrary Group

Let G be a group.

1 If H and K are subgroups of a group G , prove that $H \cap K$ is a subgroup of G . (Remember that $x \in H \cap K$ iff $x \in H$ and $x \in K$.)

2 Let H and K be subgroups of G . Prove that if $H \subseteq K$, then H is a subgroup of K .

3 By the *center* of a group G we mean the set of all the elements of G which commute with every element of G , that is,

$$C = \{a \in G : ax = xa \text{ for every } x \in G\}$$

Prove that C is a subgroup of G .

4 Let $C' = \{a \in G : (ax)^2 = (xa)^2 \text{ for every } x \in G\}$. Prove that C' is a subgroup of G .

5 Let G be a *finite* group, and let S be a nonempty subset of G . Suppose S is closed with respect to multiplication. Prove that S is a subgroup of G . (HINT: It remains to prove that S contains e and is closed with respect to inverses. Let $S = \{a_1, \dots, a_n\}$. If $a_i \in S$, consider the *distinct* elements $a_i a_1, a_i a_2, \dots, a_i a_n$.)

6 Let G be a group and $f: G \rightarrow G$ a function. A *period* of f is any element a in G such that $f(x) = f(ax)$ for every $x \in G$. Prove: The set of all the periods of f is a subgroup of G .

7 Let H be a subgroup of G , and let $K = \{x \in G : xax^{-1} \in H \text{ iff } a \in H\}$. Prove:

(a) K is a subgroup of G .

(b) H is a subgroup of K .

8 Let G and H be groups, and $G \times H$ their direct product.

(a) Prove that $\{(x, e) : x \in G\}$ is a subgroup of $G \times H$.

(b) Prove that $\{(x, x) : x \in G\}$ is a subgroup of $G \times G$.

E. Generators of Groups

1 List all the cyclic subgroups of $\langle \mathbf{z}_{10}, + \rangle$.

2 Show that $\mathbf{z}_{10}$ is generated by 2 and 5.

3 Describe the subgroup of $\mathbf{z}_{12}$ generated by 6 and 9.

4 Describe the subgroup of $\mathbf{z}$ generated by 10 and 15.

5 Show that $\mathbf{z}$ is generated by 5 and 7.

6 Show that $\mathbf{z}_2 \times \mathbf{z}_3$ is a cyclic group. Show that $\mathbf{z}_3 \times \mathbf{z}_4$ is a cyclic group.

7 Show that $\mathbf{z}_2 \times \mathbf{z}_4$ is *not* a cyclic group, but is generated by $(1, 1)$ and $(1, 2)$.

8 Suppose a group G is generated by two elements a and b . If $ab = ba$, prove that G is abelian.

F. Groups Determined by Generators and Defining Equations

1 Let G be the group $\{e, a, b, b^2, ab, ab^2\}$ whose generators satisfy $a^2 = e, b^3 = e, ba = ab^2$. Write the

table of G .

2 Let G be the group $\{e, a, b, b^2, b^3, ab, ab^2, ab^3\}$ whose generators satisfy $a^2 = e, b^4 = e, ba = ab^3$. Write the table of G . (G is called the *dihedral group* D_4 .)

3 Let G be the group $\{e, a, b, b^2, b^3, ab, ab^2, ab^3\}$ whose, generators satisfy $a^4 = e, a^2 = b^2, ba = ab^3$. Write the table of G . (G is called the *quaternion group*.)

4 Let G be the *commutative* group $\{e, a, b, c, ab, be, ac, abc\}$ whose generators satisfy $a^2 = b^2 = c^2 = e$. Write the table of G .

G. Cayley Diagrams

Every finite group may be represented by a diagram known as a Cayley diagram. A Cayley diagram consists of points joined by arrows.

There is one point for every element of the group.

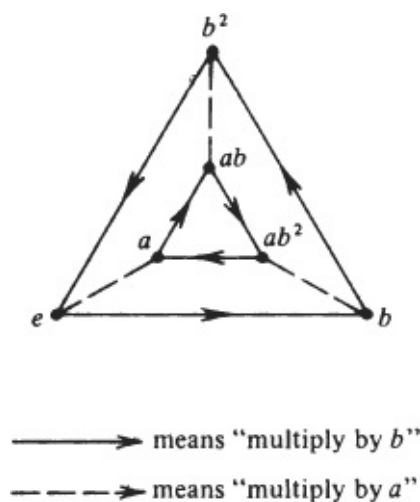
The arrows represent the result of multiplying by a generator.

For example, if G has only one generator a (that is, G is the cyclic group $\langle a \rangle$), then the arrow $\rightarrow$ represents the operation “multiply by a ”:

$$e \rightarrow a \rightarrow a^2 \rightarrow a^3 \cdots$$

If the group has two generators, say a and b , we need *two kinds of arrows*, say $\rightarrow$ and $\dashrightarrow$, where $\rightarrow$ means “multiply by a ,” and $\dashrightarrow$ means “multiply by b .”

For example, the group $G = \{e, a, b, b^2, ab, ab^2\}$ where $a^2 = e, b^3 = e$, and $ba = ab^2$ (see page 47) has the following Cayley diagram:



Moving in the *forward* direction of the arrow $\rightarrow$ means multiplying by b ,

$$x \rightarrow xb$$

whereas moving in the *backward* direction of the arrow means multiplying by b^{-1} :

$$x \leftarrow xb^{-1}$$

(Note that “multiplying x by b ” is understood to mean multiplying *on the right* by b : it means xb , not bx .) It is also a convention that if $a^2 = e$ (hence $a = a^{-1}$), then no arrowhead is used:

for if $a = a^{-1}$, then multiplying by a is the same as multiplying by a^{-1} .

The Cayley diagram of a group contains the same information as the group's table. For instance, to find the product $(ab)(ab^2)$ in the figure on page 51, we start at ab and follow the path corresponding to ab^2 (multiplying by a , then by b , then again by b), which is



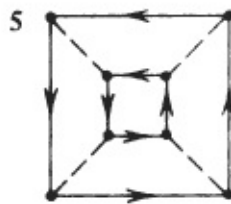
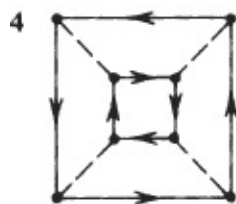
This path leads to b ; hence $(ab)(ab^2) = b$.

As another example, the inverse of ab^2 is the path which leads from ab^2 back to e . We note instantly that this is ba .

A point-and-arrow diagram is the Cayley diagram of a group iff it has the following two properties: (a) For each point x and generator a , there is exactly one a -arrow starting at x , and exactly one a -arrow ending at x ; furthermore, at most one arrow goes from x to another point y . (b) If two different paths starting at x lead to the same destination, then these two paths, starting at any point y , lead to the same destination.

Cayley diagrams are a useful way of finding new groups.

Write the table of the groups having the following Cayley diagrams: (REMARK: You may take any point to represent e , because there is perfect symmetry in a Cayley diagram. Choose e , then label the diagram and proceed.)



H. Coding Theory: Generator Matrix and Parity-Check Matrix of a Code

For the reader who does not know the subject, linear algebra will be developed in [Chapter 28](#). However, some rudiments of vector and matrix multiplication will be needed in this exercise; they are given here:

A *vector* with n components is a sequence of n numbers: $(a_1, a_2, \dots, a_n)$. The *dot product* of two vectors with n components, say the vectors $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$, is defined by

$$\mathbf{a} \cdot \mathbf{b} = (a_1, a_2, \dots, a_n) \cdot (b_1, b_2, \dots, b_n) = a_1b_1 + a_2b_2 + \dots + a_nb_n$$

that is, you multiply corresponding components and add. For example,

$$(1, 4, -2, 3) \cdot (6, 2, 4, -2) = 1(6) + 4(2) + (-2)4 + 3(-2) = 0$$

When $\mathbf{a} \cdot \mathbf{b} = 0$, as in the last example, we say that $\mathbf{a}$ and $\mathbf{b}$ are *orthogonal*.

A *matrix* is a rectangular array of numbers. An “ m by n matrix” ($m \times n$ matrix) has m rows and n columns. For example,

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & -2 & 3 \\ 4 & 1 & 1 & -3 \\ 7 & 2 & 5 & -1 \end{pmatrix}$$

is a 3×4 matrix: It has three rows and four columns. Notice that each row of $\mathbf{B}$ is a vector with four components, and each column of $\mathbf{B}$ is a vector with three components.

If $\mathbf{A}$ is any $m \times n$ matrix, let $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ be the *columns* of $\mathbf{A}$. (Each column of $\mathbf{A}$ is a vector with m components.) If $\mathbf{x}$ is any vector with m components, $\mathbf{x}\mathbf{A}$ denotes the vector

$$\mathbf{x}\mathbf{A} = (\mathbf{x} \cdot \mathbf{a}_1, \mathbf{x} \cdot \mathbf{a}_2, \dots, \mathbf{x} \cdot \mathbf{a}_n)$$

That is, the components of $\mathbf{x}\mathbf{A}$ are obtained by dot multiplying $\mathbf{x}$ by the successive columns of $\mathbf{A}$. For example, if $\mathbf{B}$ is the matrix of the previous paragraph and $\mathbf{x} = (3, 1, -2)$, then the components of $\mathbf{x}\mathbf{B}$ are

$$(3, 1, -2) \cdot (1, 4, 7) = -7$$

$$(3, 1, -2) \cdot (2, 1, 2) = 3$$

$$(3, 1, -2) \cdot (-2, 1, 5) = -15$$

$$(3, 1, -2) \cdot (3, -3, -1) = 8$$

that is, $\mathbf{x}\mathbf{B} = (-7, 3, -15, 8)$.

If $\mathbf{A}$ is an $m \times n$ matrix, let $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \dots, \mathbf{a}^{(m)}$ be the *rows* of $\mathbf{A}$. If $\mathbf{y}$ is any vector with n components, $\mathbf{A}\mathbf{y}$ denotes the vector

$$\mathbf{A}\mathbf{y} = (\mathbf{y} \cdot \mathbf{a}^{(1)}, \mathbf{y} \cdot \mathbf{a}^{(2)}, \dots, \mathbf{y} \cdot \mathbf{a}^{(m)})$$

That is, the components of $\mathbf{A}\mathbf{y}$ are obtained by dot multiplying $\mathbf{y}$ with the successive *rows* of $\mathbf{A}$. (Clearly, $\mathbf{A}\mathbf{y}$ is not the same as $\mathbf{y}\mathbf{A}$.) From linear algebra, $\mathbf{A}(\mathbf{x} + \mathbf{y}) = \mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{y}$ and $(\mathbf{A} + \mathbf{B})\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{x}$.

We shall now continue the discussion of codes begun in [Exercises F](#) and [G](#) of [Chapter 3](#). Recall that $\mathbb{B}^n$ is the set of all vectors of length n whose entries are 0s and 1s. In [Exercise F](#), page 32, it was shown that $\mathbb{B}^n$ is a group. A *code* is defined to be any subset C of $\mathbb{B}^n$. A code is called a *group code* if C is a *subgroup* of $\mathbb{B}^n$. The codes described in [Chapter 3](#), as well as all those to be mentioned in this exercise, are group codes.

An $m \times n$ matrix $\mathbf{G}$ is a *generator matrix* for the code C if C is the group generated by the rows of $\mathbf{G}$. For example, if C_1 is the code given on page 34, its generator matrix is

$$\mathbf{G}_1 = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

You may check that all eight codewords of C_1 are sums of the rows of $\mathbf{G}_1$.

Recall that every codeword consists of information digits and parity-check digits. In the code C_1 the first three digits of every codeword are information digits, and make up the *message*; the last two

digits are parity-check digits. *Encoding* a message is the process of adding the parity-check digits to the end of the message. If $\mathbf{x}$ is a message, then $E(\mathbf{x})$ denotes the encoded word. For example, recall that in C_1 the parity-check equations are $a_4 = a_1 + a_3$ and $a_5 = a_1 + a_2 + a_3$. Thus, a three-digit message $a_1a_2a_3$ is encoded as follows:

$$E(a_1, a_2, a_3) = (a_1, a_2, a_3, a_1 + a_3, a_1 + a_2 + a_3)$$

The two digits added at the end of a word are those dictated by the parity check equations. You may verify that

$$\begin{aligned} E(a_1, a_2, a_3) &= (a_1, a_2, a_3) \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \quad \begin{array}{l} \text{[Dot multiply } (a_1, a_2, a_3) \\ \text{by the successive} \\ \text{columns of } \mathbf{G}_1. \end{array} \\ &= (a_1, a_2, a_3, a_1 + a_3, a_1 + a_2 + a_3) \end{aligned}$$

This is true in all cases: If $\mathbf{G}$ is the generator matrix of a code and $\mathbf{x}$ is a message, then $E(\mathbf{x})$ is equal to the product $\mathbf{xG}$. Thus, encoding using the generator matrix is very easy: you simply multiply the message $\mathbf{x}$ by the generator matrix $\mathbf{G}$.

Now, the parity-check equations of C_1 (namely, $a_4 = a_1 + a_3$ and $a_5 = a_1 + a_2 + a_3$) can be written in the form

$$a_1 + a_3 + a_4 = 0 \quad \text{and} \quad a_1 + a_2 + a_3 + a_5 = 0$$

which is equivalent to

$$(a_1, a_2, a_3, a_4, a_5) \cdot (1, 0, 1, 1, 0) = 0$$

and

$$(a_1, a_2, a_3, a_4, a_5) \cdot (1, 1, 1, 0, 1) = 0$$

The last two equations show that a word $a_1a_2a_3a_4a_5$ is a codeword (that is, satisfies the parity-check equations) if and only if $(a_1, a_2, a_3, a_4, a_5)$ is orthogonal to both rows of the matrix:

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

$\mathbf{H}$ is called the *parity-check matrix* of the code C_1 . This conclusion may be stated as a theorem:

Theorem 1 Let $\mathbf{H}$ be the parity-check matrix of a code C in $\mathbb{B}^n$. A word $\mathbf{x}$ in $\mathbb{B}^n$ is a codeword if and only if $\mathbf{Hx} = 0$.

(Remember that $\mathbf{Hx}$ is obtained by dot multiplying x by the rows of $\mathbf{H}$.)

1 Find the generator matrix $\mathbf{G}_2$ and the parity-check matrix $\mathbf{H}_2$ of the code C_2 described in [Exercise G2](#) of [Chapter 3](#).

2 Let C_3 be the following code in $\mathbb{B}^7$: the first four positions are information positions, and the parity-check equations are $a_5 = a_2 + a_3 + a_4$, $a_6 = a_1 + a_3 + a_4$, and $a_7 = a_1 + a_2 + a_4$. (C_3 is called the *Hamming code*.) Find the generator matrix $\mathbf{G}_3$ and parity-check matrix $\mathbf{H}_3$ of C_3 .

The *weight* of a word $\mathbf{x}$ is the number of 1s in the word and is denoted by $w(\mathbf{x})$. For example, $w(11011) = 4$. The *minimum weight* of a code C is the weight of the nonzero codeword of smallest weight in the code. (See the definitions of “distance” and “minimum distance” on page 34.) Prove the following:

3 $d(\mathbf{x}, \mathbf{y}) = w(\mathbf{x} + \mathbf{y})$.

4 $w(\mathbf{x}) = d(\mathbf{x}, \mathbf{0})$, where $\mathbf{0}$ is the word whose digits are all 0s.

5 The minimum distance of a group code C is equal to the minimum weight of C .

6 (a) If $\mathbf{x}$ and $\mathbf{y}$ have even weight, so does $\mathbf{x} + \mathbf{y}$.

(b) If $\mathbf{x}$ and $\mathbf{y}$ have odd weight, $\mathbf{x} + \mathbf{y}$ has even weight.

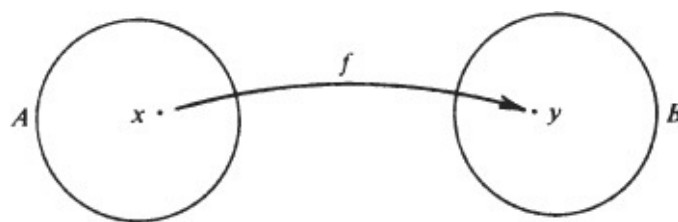
(c) If $\mathbf{x}$ has odd and $\mathbf{y}$ has even weight, then $\mathbf{x} + \mathbf{y}$ has odd weight.

7 In any group code, either all the words have even weight, or half the words have even weight and half the words have odd weight. (Use part 6 in your proof.)

8 $\mathbf{H}(\mathbf{x} + \mathbf{y}) = \mathbf{0}$ if and only if $\mathbf{H}\mathbf{x} = \mathbf{H}\mathbf{y}$, where $\mathbf{H}$ denotes the parity-check matrix of a code in $\mathbb{B}^n$ and $\mathbf{x}$ and $\mathbf{y}$ are any two words in $\mathbb{B}^n$.

FUNCTIONS

The concept of a *function* is one of the most basic mathematical ideas and enters into almost every mathematical discussion. A function is generally defined as follows: If A and B are sets, then a function from A to B is a rule which to every element x in A assigns a unique element y in B . To indicate this connection between x and y we usually write $y = f(x)$, and we call y the *image* of x under the function f .



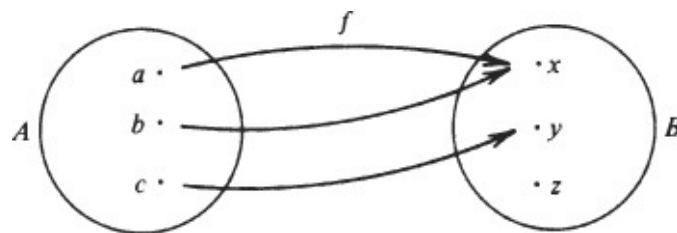
There is nothing inherently mathematical about this notion of function. For example, imagine A to be a set of married men and B to be the set of their wives. Let f be the rule which to each man assigns his wife. Then f is a perfectly good function from A to B ; under this function, each wife is the image of her husband. (No pun is intended.)

Take care, however, to note that if there were a bachelor in A then f would not qualify as a function from A to B ; for a function from A to B must assign a value in B to *every* element of A , without exception. Now, suppose the members of A and B are Ashanti, among whom polygamy is common; in the land of the Ashanti, f does not necessarily qualify as a function, for it may assign to a given member of A several wives. If f is a function from A to B , it must assign exactly *one* image to each element of A .

If f is a function from A to B it is customary to describe it by writing

$$f: A \rightarrow B$$

The set A is called the *domain* of f . The *range* of f is the subset of B which consists of *all the images of elements of A* . In the case of the function illustrated here, $\{a, b, c\}$ is the domain of f , and $\{x, y\}$ is the



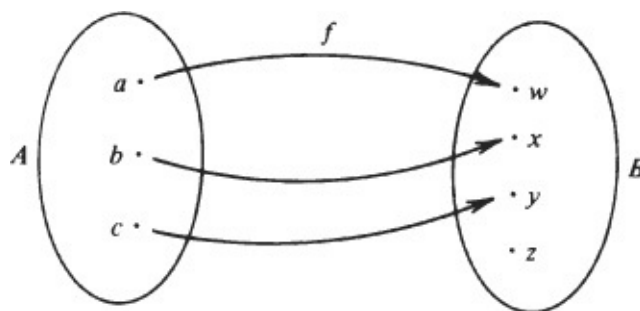
range of f (z is not in the range of f). Incidentally, this function f may be represented in the simplified notation

$$f = \begin{pmatrix} a & b & c \\ x & x & y \end{pmatrix}$$

This notation is useful whenever A is a finite set: the elements of A are listed in the top row, and beneath each element of A is its image.

It may perfectly well happen, if f is a function from A to B , that two or more elements of A *have the same image*. For example, if we look at the function immediately above, we observe that a and b both have the same image x . If f is a function for which this kind of situation *does not occur*, then f is called an *injective* function. Thus,

Definition 1 A function $f: A \rightarrow B$ is called **injective** if each element of B is the image of no more than one element of A .



The intended meaning, of course, is that each element y in B is the image of no two *distinct* elements of A . So if

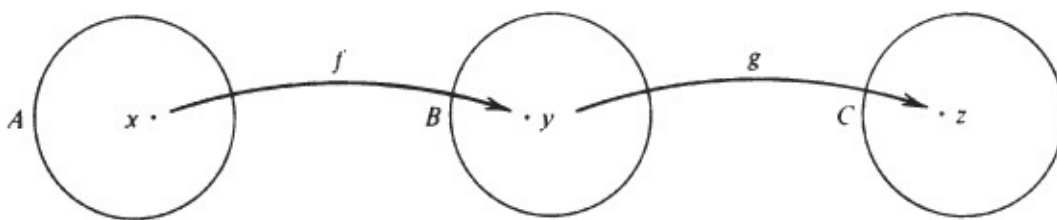
$$\begin{array}{c} x_1 \\ x_2 \end{array} \rightarrow y = f(x_1) = f(x_2)$$

that is, x_1 and x_2 have the same image y , we must require that x_1 be *equal* to x_2 . Thus, a convenient definition of “injective” is this: a function $f: A \rightarrow B$ is *injective* if and only if

$$f(x_1) = f(x_2) \quad \text{implies} \quad x_1 = x_2$$

If f is a function from A to B , there may be elements in B which are not images of elements of A . If this does *not happen*, that is, if every element of B is the image of some element of A , we say that f is *surjective*.

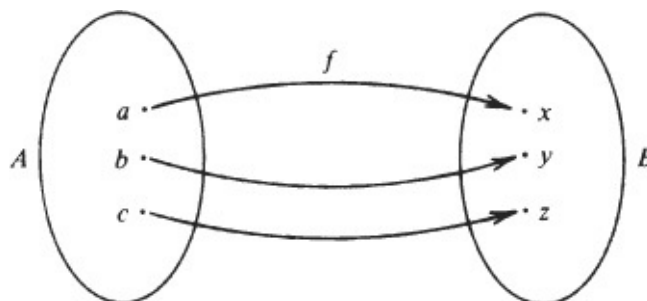
Definition 2 A function $f: A \rightarrow B$ is called **surjective** if each element of B is the image of at least one element of A .



This is the same as saying that B is the range of f .

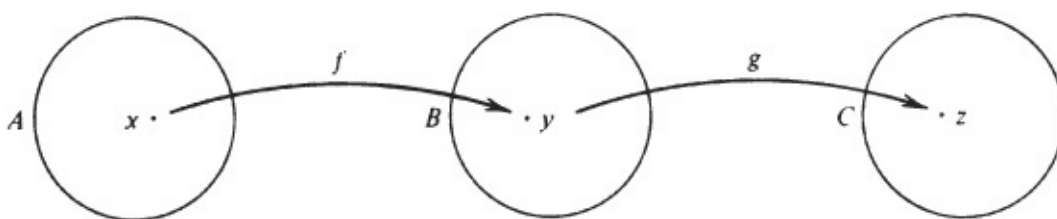
Now, suppose that f is both injective and surjective. By Definitions 1 and 2, each element of B is the image of *at least one* element of A , and *no more than one* element of A . So each element of B is the image of *exactly one* element of A . In this case, f is called a *bijective* function, or a *one-to-one correspondence*.

Definition 3 A function $f: A \rightarrow B$ is called **bijective** if it is both injective and surjective.



It is obvious that under a bijective function, each element of A has exactly one “partner” in B and each element of B has exactly one partner in A .

The most natural way of *combining* two functions is to form their “composite.” The idea is this: suppose f is a function from A to B , and g is a function from B to C . We apply f to an element x in A and get an element y in B ; then we apply g to y and get an element z in C . Thus, z is obtained from x by applying f and g in succession. The function which

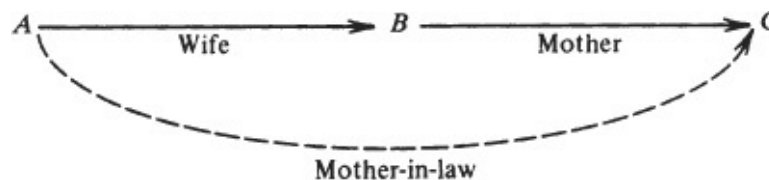


consists of *applying f and g in succession* is a function from A to C , and is called the *composite* of f and g . More precisely,

Let $f: A \rightarrow B$ and $g: B \rightarrow C$ be functions. The **composite function** denoted by $g \circ f$ is a function from A to C defined as follows:

$$[g \circ f](x) = g(f(x)) \quad \text{for every } x \in A$$

For example, consider once again a set A of married men and the set B of their wives. Let C be the set of all the mothers of members of B . Let $f: A \rightarrow B$ be the rule which to each member of A assigns his wife, and $g: B \rightarrow C$ the rule which to each woman in B assigns her mother. Then $g \circ f$ is the “mother-in-law function,” which assigns to each member of A his wife’s mother:



For another, more conventional, example, let f and g be the following functions from $\mathbb{R}$ to $\mathbb{R}$: $f(x) = 2x$; $g(x) = x + 1$. (In other words, f is the rule “multiply by 2” and g is the rule “add 1.”) Their composites are the functions $g \circ f$ and $f \circ g$ given by

$$[f \circ g](x) = f(g(x)) = 2(x + 1)$$

and

$$[g \circ f](x) = g(f(x)) = 2x + 1$$

$f \circ g$ and $g \circ f$ are different: $f \circ g$ is the rule “add 1, then multiply by 2,” whereas $g \circ f$ is the rule “multiply by 2 and then add 1.”

It is an important fact that the composite of two injective functions is injective, the composite of two surjective functions is surjective, and the composite of two bijective functions is bijective. In other words, if $f: A \rightarrow B$ and $g: B \rightarrow C$ are functions, then the following are true:

If f and g are injective, then $g \circ f$ is injective.

If f and g are surjective, then $g \circ f$ is surjective.

If f and g are bijective, then $g \circ f$ is bijective.

Let us tackle each of these claims in turn. We will suppose that f and g are injective, and prove that $g \circ f$ is injective. (That is, we will prove that if $[g \circ f](x) = [g \circ f](y)$, then $x = y$.)

Suppose $[g \circ f](x) = [g \circ f](y)$, that is,

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

Because g is injective, we get

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

and because f is injective,

$$x = y$$

Next, let us suppose that f and g are surjective, and prove that $g \circ f$ is surjective. What we need to show here is that *every* element of C is $g \circ f$ of some element of A . Well, if $z \in C$, then (because g is surjective) $z = g(y)$ for some $y \in B$; but f is surjective, so $y = f(x)$ for some $x \in A$. Thus,

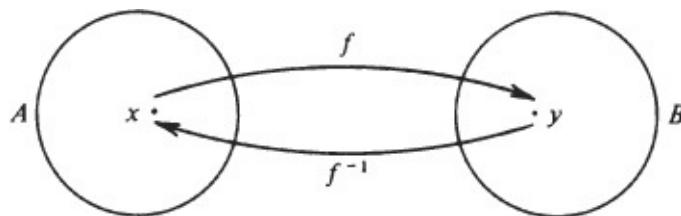
$$z = g(y) = g(f(x)) = [g \circ f](x)$$

Finally, if f and g are bijective, they are both injective and surjective. By what we have already proved, $g \circ f$ is injective and surjective, hence bijective.

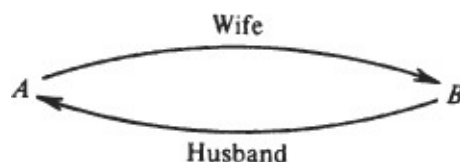
A function f from A to B may have an inverse, but it does not have to. The inverse of f , if it exists,

is a function f^{-1} (“ f inverse”) from B to A such that

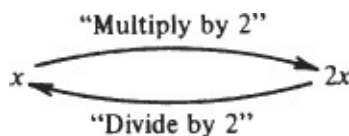
$$x = f^{-1}(y) \quad \text{if and only if} \quad y = f(x)$$



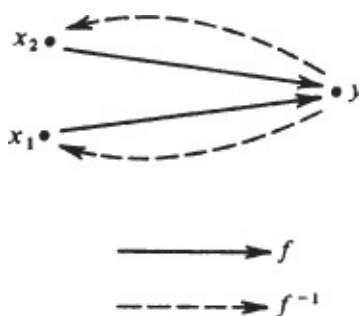
Roughly speaking, if f carries x to y then f^{-1} carries y to x . For instance, returning (for the last time) to the example of a set A of husbands and the set B of their wives, if $f: A \rightarrow B$ is the rule which to each husband assigns his wife, then $f^{-1}: B \rightarrow A$ is the rule which to each wife assigns her husband:



If we think of functions as rules, then f^{-1} is the rule which *undoes* what ever f does. For instance, if f is the real-valued function $f(x) = 2x$, then f^{-1} is the function $f^{-1}(x) = x/2$ [or, if preferred, $f^{-1}(y) = y/2$]. Indeed, the rule “divide by 2” undoes what the rule “multiply by 2” does.



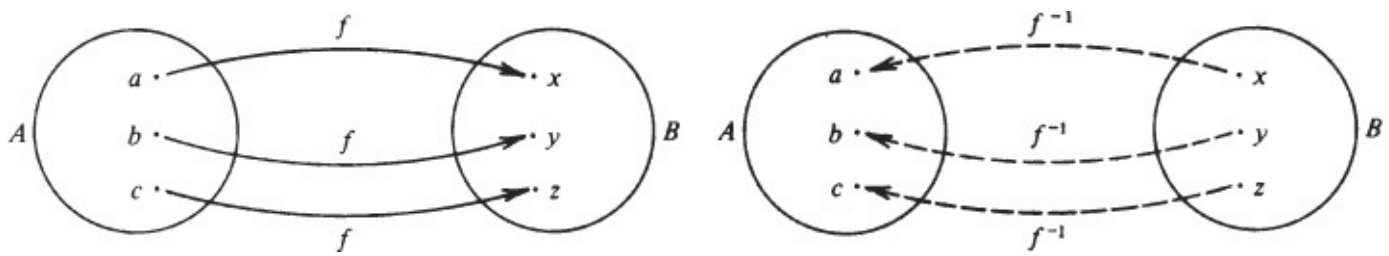
Which functions have inverses, and which others do not? If f , a function from A to B , is *not* injective, it *cannot* have an inverse; for “not injective” means there are at least two distinct elements x_1 and x_2 with the same image y ;



Clearly, $x_1 = f^{-1}(y)$ and $x_2 = f^{-1}(y)$ so $f^{-1}(y)$ is *ambiguous* (it has two different values), and this is not allowed for a function.

If f , a function from A to B is *not* surjective, there is an element y in B which is not an image of any element of A ; thus $f^{-1}(y)$ does not exist. So f^{-1} cannot be a function from B (that is, with domain B) to A .

It is therefore obvious that if f^{-1} exists, f *must be* injective and surjective, that is, bijective. On the other hand, if f is a bijective function from A to B , its inverse clearly exists and is determined by the rule that if $y = f(x)$ then $f^{-1}(y) = x$.



Furthermore, it is easy to see that the inverse of f is also a bijective function. To sum up:

A function $f: A \rightarrow B$ has an inverse if and only if it is bijective.

In that case, the inverse f^{-1} is a bijective function from B to A .

EXERCISES

A. Examples of Injective and Surjective Functions

Each of the following is a function $f: \mathbb{R} \rightarrow \mathbb{R}$. Determine

(a) whether or not f is injective, and

(b) whether or not f is surjective.

Prove your answer in either case.

Example 1 $f(x) = 2x$

f is injective.

PROOF Suppose $f(a) = f(b)$, that is,

$$2a = 2b$$

Then

$$a = b$$

Therefore f is injective. ■

f is surjective.

PROOF Take any element $y \in \mathbb{R}$. Then $y = 2(y/2) = f(y/2)$.

Thus, every $y \in \mathbb{R}$ is equal to $f(x)$ for $x = y/2$.

Therefore f is surjective. ■

Example 2 $f(x) = x^2$

f is not injective.

PROOF By exhibiting a counterexample: $f(2) = 4 = f(-2)$, although $2 \neq -2$. ■

f is not surjective.

PROOF By exhibiting a counterexample: -1 is not equal to $f(x)$ for any $x \in \mathbb{R}$. ■

1 $f(x) = 3x + 4$

2 $f(x) = x^3 + 1$

3 $f(x) = |x|$

4 $f(x) = x^3 - 3x$

5 $f(x) = \begin{cases} x & \text{if } x \text{ is rational} \\ 2x & \text{if } x \text{ is irrational} \end{cases}$

6 $f(x) = \begin{cases} 2x & \text{if } x \text{ is an integer} \\ x & \text{otherwise} \end{cases}$

7 Determine the range of each of the functions in parts 1 to 6.

B. Functions on $\mathbb{R}$ and $\mathbb{Z}$

Determine whether each of the functions listed in parts 1–4 is or is not (a) injective and (b) surjective. Proceed as in [Exercise A](#).

1 $f: \mathbb{R} \rightarrow (0, \infty)$, defined by $f(x) = e^x$.

2 $f: (0,1) \rightarrow \mathbb{R}$, defined by $f(x) = \tan x$.

3 $f: \mathbb{Z} \rightarrow \mathbb{Z}$, defined by $f(x) =$ the least integer greater than or equal to x .

4 $f: \mathbb{Z} \rightarrow \mathbb{Z}$, defined by $f(n) = \begin{cases} n+1 & \text{if } n \text{ is even} \\ n-1 & \text{if } n \text{ is odd} \end{cases}$

5 Find a bijective function f from the set $\mathbb{Z}$ of the integers to the set E of the *even* integers.

C. Functions on Arbitrary Sets and Groups

Determine whether each of the following functions is or is not (a) injective and (b) surjective. Proceed as in [Exercise A](#).

In parts 1 to 3, A and B are sets, and $A \times B$ denotes the set of all the ordered pairs (x,y) as x ranges over A and y over B .

1 $f: A \times B \rightarrow A$, defined by $f(x,y) = x$.

2 $f: A \times B \rightarrow B \times A$, defined by $f(x,y) = (y,x)$.

3 $f: A \times B \rightarrow B$, defined by $f(x,b) = b$, where b is a fixed element of B .

4 G is a group, $a \in G$, and $f: G \rightarrow G$ is defined by $f(x) = ax$.

5 G is a group and $f: G \rightarrow G$ is defined by $f(x) = x^{-1}$.

6 G is a group and $f: G \rightarrow G$ is defined by $f(x) = x^2$.

D. Composite Functions

In parts 1–3 find the composite function, as indicated.

1 $f: \mathbb{R} \rightarrow \mathbb{R}$ is defined by $f(x) = \sin x$.

$g: \mathbb{R} \rightarrow \mathbb{R}$ is defined by $g(x) = e^x$.

Find $f \circ g$ and $g \circ f$.

2 A and B are sets; $f: A \times B \rightarrow B \times A$ is given by $f(x,y) = (y,x)$.

$g: B \times A \rightarrow B$ is given by $g(y,x) = y$.

Find $g \circ f$.

3 $f: (0,1) \rightarrow \mathbb{R}$ is defined by $f(x) = 1/x$.

$g: \mathbb{R} \rightarrow \mathbb{R}$ is defined by $g(x) = \ln x$.

Find $g \circ f$. Explain why $f \circ g$ is undefined.

4 In school, Jack and Sam exchanged notes in a code f which consists of spelling every word backwards and interchanging every letter s with t . Alternatively, they use a code g which interchanges the letters a with o , i with u , e with y , and s with t . Describe the codes $f \circ g$ and $g \circ f$. Are they the same?

5 $A = \{a, b, c, d\}$; f and g are functions from A to A ; in the tabular form described on page 57, they are given by

$$f = \begin{pmatrix} a & b & c & d \\ a & c & a & c \end{pmatrix} \quad g = \begin{pmatrix} a & b & c & d \\ b & a & b & a \end{pmatrix}$$

Give $f \circ g$ and $g \circ f$ in the same tabular form.

6 G is a group, and a and b are elements of G .

$f: G \rightarrow G$ is defined by $f(x) = ax$.

$g: G \rightarrow G$ is defined by $g(x) = bx$.

Find $f \circ g$ and $g \circ f$.

7 Indicate the domain and range of each of the composite functions you found in parts 1 to 6.

E. Inverses of Functions

Each of the following functions f is bijective. Describe its inverse.

1 $f: (0, \infty) \rightarrow (0, \infty)$, defined by $f(x) = 1/x$.

2 $f: \mathbb{R} \rightarrow (0, \infty)$, defined by $f(x) = e^x$.

3 $f: \mathbb{R} \rightarrow \mathbb{R}$, defined by $f(x) = x^3 + 1$.

4 $f: \mathbb{R} \rightarrow \mathbb{R}$, defined by $f(x) = \begin{cases} 2x & \text{if } x \text{ is rational} \\ 3x & \text{if } x \text{ is irrational} \end{cases}$

5 $A = \{a, b, c, d\}$, $B = \{1, 2, 3, 4\}$ and $f: A \rightarrow B$ is given by

$$f = \begin{pmatrix} a & b & c & d \\ 3 & 1 & 2 & 4 \end{pmatrix}$$

6 G is a group, $a \in G$, and $f: G \rightarrow G$ is defined by $f(x) = ax$.

F. Functions on Finite Sets

1 The members of the U.N. Peace Committee must choose, from among themselves, a presiding officer of their committee. For each member x , let $f(x)$ designate that member's choice for officer. If no two members vote alike, what is the range of f ?

2 Let A be a finite set. Explain why any injective function $f: A \rightarrow A$ is necessarily surjective. (Look at part 1.)

3 If A is a finite set, explain why any surjective function $f: A \rightarrow A$ is necessarily injective.

4 Are the statements in parts 2 and 3 true when A is an infinite set? If not, give a counterexample.

5 If A has n elements, how many functions are there from A to A ? How many bijective functions are there from A to A ?

G. Some General Properties of Functions

In parts 1 to 3, let A , B , and C be sets, and let $f: A \rightarrow B$ and $g: B \rightarrow C$ be functions.

1 Prove that if $g \circ f$ is injective, then f is injective.

- 2 Prove that if $g \circ f$ is surjective, then g is surjective.
- 3 Parts 1 and 2, together, tell us that if $g \circ f$ is bijective, then f is injective and g is surjective. Is the converse of this statement true: If $xy0$ is injective and g surjective, is $g \circ f$ bijective? (If “yes,” prove it; if “no,” give a counterexample.)
- 4 Let $f: A \rightarrow B$ and $g: B \rightarrow A$ be functions. Suppose that $y = f(x)$ iff $x = g(y)$. Prove that f is bijective, and $g = f^{-1}$

H. Theory of Automata

Digital computers and other electronic systems are made up of certain basic control circuits. Underlying such circuits is a fundamental mathematical notion, the notion of *finite automata*, also known as *finite-state machines*.

A finite automaton receives information which consists of sequences of symbols from some *alphabet* A . A typical input sequence is a word $x = x_1x_2 \dots x_n$, where $x_1 x_2, \dots$ are symbols in the alphabet A . The machine has a set of internal components whose combined state is called the *internal state* of the machine. At each time interval the machine “reads” one symbol of the incoming input sequence and responds by going into a new internal state: the response depends both on the symbol being read and on the machine’s present internal state. Let S denote the set of internal states; we may describe a particular machine by specifying a function $\alpha: S \times A \rightarrow S$. If s_i is an internal state and a_j is the symbol currently being read, then $\alpha(s_i a_j) = s_k$ is the machine’s next state. (That is, the machine, while in state s_i responds to the symbol a_j by going into the new state s_k .) The function α is called the *next-state function* of the machine.

Example Let M_1 be the machine whose alphabet is $A = \{0,1\}$, whose set of internal states is $S = \{s_0, s_1\}$, and whose next-state function is given by the table

Present state	0	1	← Input
s_0	s_0	s_1	
s_1	s_1	s_0	

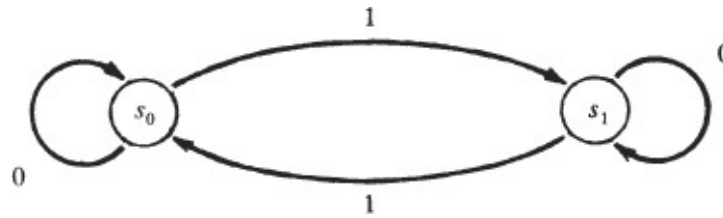
(The table asserts: When in state s_0 and reading 0, remain in s_0 . When in s_0 and reading 1, go to state s_1 . When in s_1 and reading 0, remain in s_1 . When in s_1 and reading 1, go to state s_0 .)

A possible use of M_1 is as a *parity-check machine*, which may be used in decoding information arriving on a communication channel. Assume the incoming information consists of sequences of five symbols, 0s and 1s, such as 10111. The machine starts off in state s_0 . It reads the first digit, 1, and as dictated by the table above, goes into state s_1 . Then it reads the second digit, 0, and remains in s_1 . When it reads the third digit, 1, it goes into state s_0 , and when it reads the fourth digit, 1, it goes into s_1 . Finally, it reaches the last digit, which is the parity-check digit: if the sum of the first four digits is even, the parity-check digit is 0; otherwise it is 1. When the parity-check digit, 1, is read, the machine goes into state s_0 . Ending in state s_0 indicates that the parity-check digit is correct. If the machine ends in s_1 , this indicates that the parity-check digit is incorrect and there has been an error of transmission.

A machine can also be described with the aid of a *state diagram*, which consists of circles interconnected by arrows: the notation



means that if the machine is in state s_i when x is read, it goes into state s_j . The diagram of the machine M_1 of the previous example is



In parts 1–4 describe the machines which are able to carry out the indicated functions. For each machine, give the alphabet A , the set of states S , and the table of the next-state function. Then draw the state diagram of the machine.

1 The input alphabet consists of four letters, a , b , c , and d . For each incoming sequence, the machine determines whether that sequence contains exactly three a 's.

2 The same conditions pertain as in part 1, but the machine determines whether the sequence contains *at least* three a 's.

3 The input alphabet consists of the digits 0, 1, 2, 3, 4. The machine adds the digits of an input sequence; the addition is modulo 5 (see page 27). The sum is to be given by the machine's state after the last digit is read.

4 The machine tells whether or not an incoming sequence of 0's and 1's ends with 111.

5 If M is a machine whose next-state function is α , define $\bar{\alpha}$ as follows: If $\mathbf{x}$ is an input sequence and the machine (in state s_i .) begins reading $\mathbf{x}$, then $\bar{\alpha}(s_i, \mathbf{x})$ is the state of the machine after the last symbol of $\mathbf{x}$ is read. For instance, if M_1 is the machine of the example given above, then $\bar{\alpha}(s_0, 11010) = s_1$. (The machine is in s_0 before the first symbol is read; each 1 alters the state, but the 0s do not. Thus, after the last 0 is read, the machine is in state s_1 .)

(a) For the machine M_1 , give $\bar{\alpha}(s_0, \mathbf{x})$ for all three-digit sequences $\mathbf{x}$.

(b) For the machine of part 1, give $\bar{\alpha}(s_i, \mathbf{x})$ for each state s_i and every two-letter sequence $\mathbf{x}$.

6 With each input sequence $\mathbf{x}$ we associate a function $T_{\mathbf{x}}: S \rightarrow S$ called a *state transition function*, defined as follows:

$$T_{\mathbf{x}}(s_i) = \bar{\alpha}(s_i, \mathbf{x})$$

For the machine M_1 of the example, if $\mathbf{x} = 11010$, $T_{\mathbf{x}}$ is given by

$$T_{\mathbf{x}}(s_0) = s_1 \quad \text{and} \quad T_{\mathbf{x}}(s_1) = s_0$$

(a) Describe the transition function $T_{\mathbf{x}}$ for the machine M_1 and the following sequences: $\mathbf{x} = 01001$, $\mathbf{x} = 10011$, $\mathbf{x} = 01010$.

(b) Explain why M_1 has only two *distinct* transition functions. [Note: Two functions f and g are equal if $f(x) = g(x)$ for every x ; otherwise they are distinct.]

(c) For the machine of part 1, describe the transition function T_1 for the following $\mathbf{x}$: $\mathbf{x} = abbca$, $\mathbf{x} =$

$babac, \mathbf{x} = ccbaa$.

(d) How many *distinct* transition functions are there for the machine of part 3?

I. Automata, Semigroups, and Groups

By a *semigroup* we mean a set A with an associative operation. (There does not need to be an identity element, nor do elements necessarily have inverses.) Every group is a semigroup, though the converse is clearly false. With every semigroup A we associate an automaton $M = M(A)$ called the *automaton of the semigroup* A . The alphabet of M is A , the set of states also is A , and the next-state function is $\alpha(s, a) = sa$ [or $\alpha(s, a) = s + a$ if the operation of the semigroup is denoted additively].

1 Describe $M(\mathbb{Z}_4)$. That is, give the table of its next-state function, as well as its state diagram.

2 Describe $M(S_3)$.

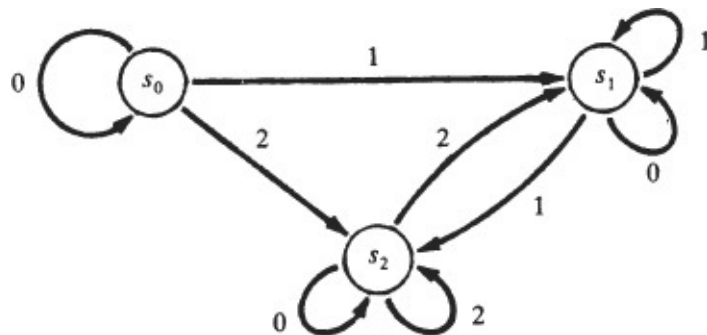
If M is a machine and S is the set of states of M , the *state transition functions* of M (defined in [Exercise H6](#) of this chapter) are functions from S to S . In the next exercise you will be asked to show that $T_y \circ T_x = T_{xy}$; that is, the composite of two transition functions is a transition function. Since the composition of functions is associative [$f \circ (g \circ h) = (f \circ g) \circ h$], it follows that the set of all transition functions, with the operation $\circ$, is a *semigroup*. It is denoted by $\mathcal{S}(M)$ and called the *semigroup of the machine* M .

3 Prove that $T_{xy} = T_y \circ T_x$.

4 Let M_1 be the machine of the example in [Exercise H](#) above. Give the table of the semigroup $\mathcal{S}(M_1)$. Does $\mathcal{S}(M_1)$ have an identity element? Is $\mathcal{S}(M_1)$ a group?

5 Let M_2 be the machine of [Exercise H3](#). How many *distinct* functions are there in $\mathcal{S}(M_2)$? Give the table of $\mathcal{S}(M_2)$. Is $\mathcal{S}(M_2)$ a group? (Why?)

6 Find the table of $\mathcal{S}(M)$ if M is the machine whose state diagram is



PARTITIONS AND EQUIVALENCE RELATIONS

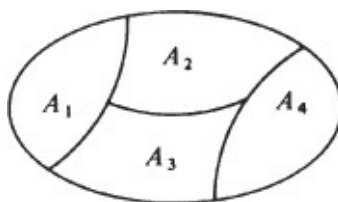
Imagine emptying a jar of coins onto a table and sorting them into separate piles, one with the pennies, one with the nickels, one with the dimes, one with the quarters, and one with the half-dollars. This is a simple example of *partitioning* a set. Each separate pile is called a class of the partition; the jarful of coins has been partitioned into five classes.

Here are some other examples of partitions: The distributor of farm-fresh eggs usually sorts the daily supply according to size, and separates the eggs into three classes called “large,” “medium,” and “small.”

The delegates to the Democratic national convention may be classified according to their home state, thus falling into 50 separate classes, one for each state.

A student files class notes according to subject matter; the notebook pages are separated into four distinct categories, marked (let us say) “algebra,” “psychology,” “English,” and “American history.”

Every time we file, sort, or classify, we are performing the simple act of partitioning a set. To partition a set A is to separate the elements of A into nonempty subsets, say A_1, A_2, A_3 etc., which are called the *classes* of the partition. Any two distinct classes, say A_i and A_j are *disjoint*, which means they have no elements in common. And the union of the classes is all of A .



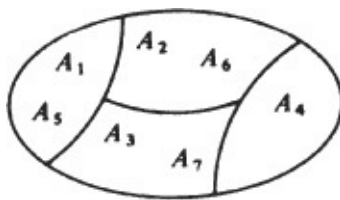
Instead of dealing with the *process* of partitioning a set (which is awkward mathematically), it is more convenient to deal with the *result* of partitioning a set. Thus, $\{A_1, A_2, A_3, A_4\}$, in the illustration above, is called a *partition* of A . We therefore have the following definition:

A partition of a set A is a family $\{A_i; i \in I\}$ of nonempty subsets of A which are mutually disjoint and whose union is all of A .

The notation $\{A_i; i \in I\}$ is the customary way of representing a family of sets $\{A_i, A_j, A_k, \dots\}$ consisting of one set A_i for each index i in I . (The elements of I are called *indices*, the notation $\{A_i; i \in$

$I\}$ may be read: the family of sets A_i , as i ranges over I .)

Let $\{A_i: i \in I\}$ be a partition of the set A . We may think of the indices $i, j, k, \dots$ as labels for naming the classes $A_i, A_j, A_k, \dots$. Now, in practical problems, it is very inconvenient to insist that each class be named once and only once. It is simpler to allow repetition of indexing whenever convenience dictates. For example, the partition illustrated previously might also be represented like this, where A_1 is the same class as A_5 , A_2 is the same as A_6 , and A_3 is the same as A_7 .



As we have seen, any two *distinct* classes of a partition are disjoint; this is the same as saying that *if two classes are not disjoint, they must be equal*. The other condition for the classes of a partition of A is that their union must be equal to A ; this is the same as saying that *every element of A lies in one of the classes*. Thus, we have the following, more explicit definition of partition:

By a **partition** of a set A we mean a family $\{A_i: i \in I\}$ of nonempty subsets of A such that

- (i) If any two classes, say A_i and A_j , have a common element x (that is, are not disjoint), then $A_i = A_j$ and
- (ii) Every element x of A lies in one of the classes.

We now turn to another elementary concept of mathematics. A *relation* on a set A is any statement which is either true or false for each ordered pair (x, y) of elements of A . Examples of relations, on appropriate sets, are “ $x = y$,” “ $x < y$,” “ x is parallel to y ,” “ x is the offspring of y ,” and so on. An especially important kind of relation on sets is an *equivalence relation*. Such a relation, will usually be represented by the symbol $\sim$, so that $x \sim y$ may be read “ x is equivalent to y .” Here is how equivalence relations are defined:

By an equivalence relation on a set A we mean a relation $\sim$ which is

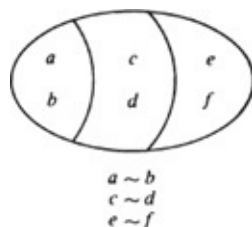
Reflexive: that is, $x \sim x$ for every x in A ;

Symmetric: that is, if $x \sim y$, then $y \sim x$; and

Transitive: that is, if $x \sim y$ and $y \sim z$, then $x \sim z$.

The most obvious example of an equivalence relation is equality, but there are many other examples, as we shall be seeing soon. Some examples from our everyday experience are “ x weighs the same as y ,” “ x is the same color as y ,” “ x is synonymous with y ,” and so on.

Equivalence relations also arise in a natural way out of partitions. Indeed, if $\{A_i: i \in I\}$ is a partition of A , we may define an equivalence relation $\sim$ on A by letting $x \sim y$ iff x and y are in the same class of the partition.



In other words, we call two elements “equivalent” if they are members of the same class. It is easy to

see that the relation $\sim$ is an equivalence relation on A . Indeed, $x \sim x$ because x is in the same class as x ; next, if x and y are in the same class, then y and x are in the same class; finally, if x and y are in the same class, and y and z are in the same class, then x and z are in the same class. Thus, $\sim$ is an equivalence relation on A ; it is called the *equivalence relation determined by the partition* $\{A_i: i \in I\}$.

Let $\sim$ be an equivalence relation on A and x an element of A . The set of all the elements equivalent to x is called the *equivalence class* of x , and is denoted by $[x]$. In symbols,

$$[x] = \{y \in A: y \sim x\}$$

A useful property of equivalence classes is this:

Lemma *If $x \sim y$, then $[x] = [y]$.*

In other words, if two elements are equivalent, they have the same equivalence class. The proof of this lemma is fairly obvious, for if $x \sim y$, then the elements equivalent to x are the same as the elements equivalent to y .

For example, let us return to the jarful of coins we discussed earlier. If A is the set of coins in the jar, call any two coins “equivalent” if they have the same value: thus, pennies are equivalent to pennies, nickels are equivalent to nickels, and so on. If x is a particular nickel in A , then $[x]$, the equivalence class of x , is the class of all the nickels in A . If y is a particular dime, then $[y]$ is the pile of all the dimes; and so forth. There are exactly five distinct equivalence classes. If we apply the lemma to this example, it states simply that if two coins are equivalent (that is, have the same value), they are in the same pile. By the way, the five equivalence classes obviously form a partition of A ; this observation is expressed in the next theorem.

Theorem *If $\sim$ is an equivalence relation on A , the family of all the equivalence classes, that is, $\{[x]: x \in A\}$, is a partition of A .*

This theorem states that if $\sim$ is an equivalence relation on A and we sort the elements of A into distinct classes by placing each element with the ones equivalent to it, we get a partition of A .

To prove the theorem, we observe first that each equivalence class is a nonempty subset of A . (It is nonempty because $x \sim x$, so $x \in [x]$.) Next, we need to show that any two distinct classes are disjoint—or, equivalently, that if two classes $[x]$ and $[y]$ have a common element, they are equal. Well, if $[x]$ and $[y]$ have a common element u , then $u \sim x$ and $u \sim y$. By the symmetric and transitive laws, $x \sim y$. Thus, $[x] = [y]$ by the lemma.

Finally, we must show that every element of A lies in some equivalence class. This is true because $x \in [x]$. Thus, the family of all the equivalence classes is a partition of A .

When $\sim$ is an equivalence relation on A and A is partitioned into its equivalence classes, we call this partition the *partition determined by the equivalence relation* $\sim$.

The student may have noticed by now that the two concepts of *partition* and *equivalence relation*, while superficially different, are actually twin aspects of the same structure on sets. Starting with an equivalence relation on A , we may partition A into equivalence classes, thus getting a partition of A . But from this partition we may retrieve the equivalence relation, for any two elements x and y are equivalent iff they lie in the same class of the partition.

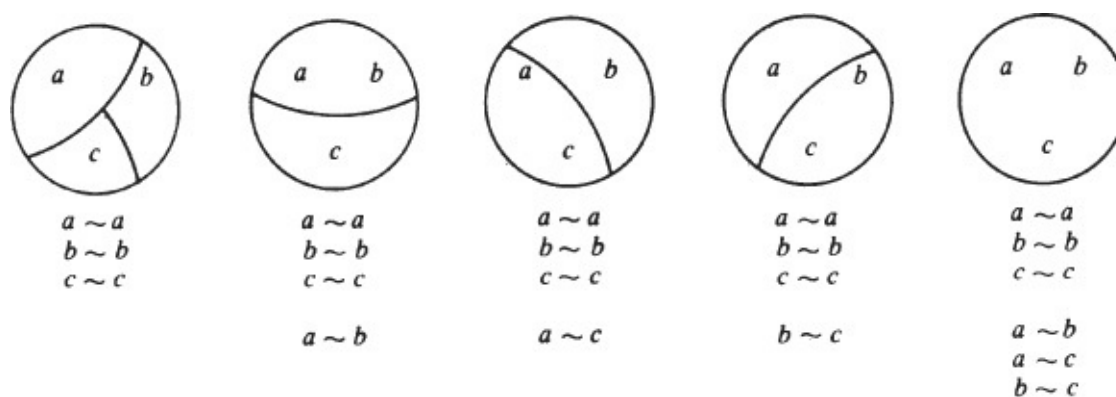
Going the other way, we may begin with a partition of A and *define* an equivalence relation by letting any two elements be equivalent iff they lie in the same class. We may then retrieve the partition by partitioning A into equivalence classes.

As a final example, let A be a set of poker chips of various colors, say red, blue, green, and white. Call any two chips “equivalent” if they are the same color. This equivalence relation has four

equivalence classes: the set of all the red chips, the set of blue chips, the set of green chips, and the set of white chips. These four equivalence classes are a partition of A .

Conversely, if we begin by partitioning the set A of poker chips into four classes according to their color, this partition determines an equivalence relation whereby chips are equivalent iff they belong to the same class. This is precisely the equivalence relation we had previously.

A final comment is in order. In general, there are many ways of partitioning a given set A ; each partition determines (and is determined by) *exactly one specific equivalence relation* on A . Thus, if A is a set of three elements, say a , b , and c , there are five ways of partitioning A , as indicated by the accompanying illustration. Under each partition is written the equivalence relation determined by that partition.



Once again, each partition of A determines, and is determined by, exactly one equivalence relation on A .

EXERCISES

A. Examples of Partitions

Prove that each of the following is a partition of the indicated set. Then describe the equivalence relation associated with that partition.

1 For each integer $r \in \{0, 1, 2, 3, 4\}$, let A_r be the set of all the integers which leave a remainder of r when divided by 5. (That is, $x \in A_r$ iff $x = 5q + r$ for some integer q .) Prove: $\{A_0, A_1, A_2, A_3, A_4\}$ is a partition of $\mathbb{Z}$.

2 For each integer n , let $A_n = \{x \in \mathbb{Q} : n \leq x < n + 1\}$. Prove $\{A_n : n \in \mathbb{Z}\}$ is a partition of $\mathbb{Q}$.

3 For each rational number r , let $A_r = \{(m, n) \in \mathbb{Z} \times \mathbb{Z}^* : m/n = r\}$. Prove that $\{A_r : r \in \mathbb{Q}\}$ is a partition of $\mathbb{Z} \times \mathbb{Z}^*$, where $\mathbb{Z}^*$ is the set of nonzero integers.

4 For $r \in \{0, 1, 2, \dots, 9\}$, let A_r be the set of all the integers whose units digit (in decimal notation) is equal to r . Prove: $\{A_0, A_1, A_2, \dots, A_9\}$ is a partition of $\mathbb{Z}$.

5 For any rational number x , we can write $x = q + n/m$ where q is an integer and $0 \leq n/m < 1$. Call n/m the *fractional part* of x . For each rational $r \in \{x : 0 \leq x < 1\}$, let $A_r = \{x \in \mathbb{Q} : \text{the fractional part of } x \text{ is equal to } r\}$. Prove: $\{A_r : 0 \leq r < 1\}$ is a partition of $\mathbb{Q}$.

6 For each $r \in \mathbb{R}$, let $A_r = \{(x, y) \in \mathbb{R} \times \mathbb{R} : x - y = r\}$. Prove: $\{A_r : r \in \mathbb{R}\}$ is a partition of $\mathbb{R} \times \mathbb{R}$.

B. Examples of Equivalence Relations

Prove that each of the following is an equivalence relation on the indicated set. Then describe the

partition associated with that equivalence relation.

- 1 In $\mathbf{Z}$, $m \sim n$ iff $|m| = |n|$.
- 2 In $\mathbf{Q}$, $r \sim s$ iff $r - s \in \mathbf{Z}$.
- 3 Let $[x]$ denote the greatest integer $\leq x$. In $\mathbf{R}$, let $a \sim b$ iff $[a] = [b]$.
- 4 In $\mathbf{Z}$, let $m \sim n$ iff $m - n$ is a multiple of 10.
- 5 In $\mathbf{R}$, let $a \sim b$ iff $a - b \in \mathbf{Q}$.
- 6 In $\mathcal{F}(\mathbf{R})$, let $f \sim g$ iff $f(0) = g(0)$.
- 7 In $\mathcal{F}(\mathbf{R})$, let $f \sim g$ iff $f(x) = g(x)$ for all $x > c$, where c is some fixed real number.
- 8 If C is any set, P_C denotes the set of all the subsets of C . Let $D \subseteq C$. In P_C , let $A \sim B$ iff $A \cap D = B \cap D$.
- 9 In $\mathbf{R} \times \mathbf{R}$, let $(a, b) \sim (c, d)$ iff $a^2 + b^2 = c^2 + d^2$.
- 10 In $\mathbf{R}^*$, let $a \sim b$ iff $a/b \in \mathbf{Q}$, where $\mathbf{R}^*$ is the set of nonzero real numbers.

C. Equivalence Relations and Partitions of $\mathbf{R} \times \mathbf{R}$

In parts 1 to 3, $\{A_r: r \in \mathbf{R}\}$ is a family of subsets of $\mathbf{R} \times \mathbf{R}$. Prove it is a partition, describe the partition geometrically, and give the corresponding equivalence relation.

- 1 For each $r \in \mathbf{R}$, $A_r = \{(x, y): y = 2x + r\}$.
- 2 For each $r \in \mathbf{R}$, $A_r = \{(x, y): x^2 + y^2 = r^2\}$.
- 3 For each $r \in \mathbf{R}$, $A_r = \{(x, y): y = |x| + r\}$.

In parts 4 to 6, an equivalence relation on $\mathbf{R} \times \mathbf{R}$ is given. Prove it is an equivalence relation, describe it geometrically, and give the corresponding partition.

- 4 $(x, y) \sim (u, v)$ iff $ax^2 + by^2 = au^2 + bv^2$ (where $a, b > 0$).
- 5 $(x, y) \sim (u, v)$ iff $x + y = u + v$
- 6 $(x, y) \sim (u, v)$ iff $x^2 - y = u^2 - v$.

D. Equivalence Relations on Groups

Let G be a group. In each of the following, a relation on G is defined. Prove it is an equivalence relation. Then describe the equivalence class of e .

- 1 If H is a subgroup of G , let $a \sim b$ iff $ab^{-1} \in H$
- 2 If H is a subgroup of G , let $a \sim b$ iff $a^{-x}b \in H$. Is this the same equivalence relation as in part 1? Prove, or find a counterexample.
- 3 Let $a \sim b$ iff there is an $x \in G$ such that $a = xbx^{-1}$.
- 4 Let $a \sim b$ iff there is an integer k such that $a^k = b^k$
- # 5 Let $a \sim b$ iff ab^{-1} commutes with every $x \in G$.
- 6 Let $a \sim b$ iff ab^{-1} is a power of c (where c is a fixed element of G).

E. General Properties of Equivalence Relations and Partitions

- 1 Let $\{A_i: i \in I\}$ be a partition of A . Let $\{B_j: j \in J\}$ be a partition of B . Prove that $\{A_i \times B_j, (i, j) \in I \times J\}$ is a partition of $A \times B$.

- 2** Let $\sim_I$ be the equivalence relation corresponding to the above partition of A , and let $\sim_J$ be the equivalence relation corresponding to the partition of B . Describe the equivalence relation corresponding to the above partition of $A \times B$.
- 3** Let $f: A \rightarrow B$ be a function. Define $\sim$ by $a \sim b$ iff $f(a) = f(b)$. Prove that $\sim$ is an equivalence relation on A . Describe its equivalence classes.
- 4** Let $f: A \rightarrow B$ be a function, and let $\{B_i: i \in I\}$ be a partition of B . Prove that $\{f^{-1}(B_i): i \in I\}$ is a partition of A . If $\sim_I$ is the equivalence relation corresponding to the partition of B , describe the equivalence relation corresponding to the partition of A . [REMARK: For any $C \subseteq B$, $f^{-1}(C) = \{x \in A: f(x) \in C\}$.]
- 5** Let $\sim_1$ and $\sim_2$ be distinct equivalence relations on A . Define $\sim_3$ by $a \sim_3 b$ iff $a \sim_1 b$ and $a \sim_2 b$. Prove that $\sim_3$ is an equivalence relation on A . If $[x]$, denotes the equivalence class of x for $\sim_i$ ($i = 1, 2, 3$), prove that $[x]_3 = [x]_1 \cap [x]_2$.