BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
WORK INTEGRATED LEARNING PROGRAMMES

COURSE HANDOUT

Part A: Content Design

Course Title	Mathematical Foundations for Data Science
Course No(s)	
Credit Units	4
Course Author	G Venkiteswaran
Version No	2
Date	15.09.2019

Course Description

Vector and matrix algebra, systems of linear algebraic equations and their solutions; eigenvalues, eigenvectors and diagonalization of matrices; graphs and digraphs; trees, lists and their uses; partially ordered sets and lattices; Boolean algebras and Boolean expressions;

Course Objectives

No	Objective- The course aims to
CO1	Introduce concepts in linear algebra and to use it as a platform to model physical problems.
CO2	Provide techniques for analytical and numerical solutions of linear equations and introduce the concept of convergence.
CO3	Utilize concepts of linear algebra and calculus in solving optimization problems.
CO4	Introduce some of the mathematical structures, concepts and notations used in discrete mathematics.
CO5	Introduce some concepts from graph theory, partially ordered sets, Boolean algebras.

Text Book(s)

No	**Author(s), Title, Edition, Publishing House**
T1	Erwin Kreyszig, Advanced Engineering Mathematics, Wiley India, 9 th Edition, 2011
T2	Kenneth H. Rosen, Discrete Mathematics and its Applications, Tata McGraw Hill, 7th Ed., 2011.

Reference Book(s) & other resources

No	**Author(s), Title, Edition, Publishing House**
R1	K Hoffman and R Kunze, Linear Algebra, Pearson Education, 2nd Edition, 2005.
R2	Kolman, Busby, Ross and Rehman, Discrete Mathematical Structures for Computer Science,

	Pearson Education, 6th Edition, 2017

Content Structure

No	Title of the module	References
M1	1. Matrices, System of equations, determinants and inverse of a matrix 1.1. Matrix Algebra-Row-reduced echelon form of a matrix, inverse of a matrix 1.2. System of linear equations, Consistency and inconsistency of system of linear equations	T1: Sec 7.1 – 7.3, 7.5, 7.8
M2	2. Vector spaces and Linear transformations 2.1 Vector space, subspace and span of a set, Linear dependence and independence of a set of vectors, basis and dimension 2.2. Linear transformation, rank and nullity	T1: Sec 7.4, 7.9, R1: Sec 3.2
M3	3. Eigenvalues, Eigenvectors and singular values 3.1. Eigenvalues 3.2. Eigenvectors 3.3. Singular value decomposition	T1: Sec 8.2, 8.3 and class notes
M4	4. Numerical linear algebra 4.1. Gauss elimination with partial pivoting and scaling 4.2. Iterative methods for solving linear system of equations	T1: Sec 20.1
M5	5. Matrix Eigenvalue Problems 5.1. Eigenvalue problems in linear system of equations 5.2. Power method for finding the dominant eigenvalue	T1: Sec 20.3, 20.8
M6	6. Linear and non-linear optimization 6.1 Basics of calculus 6.2 Linear optimization using simplex method and sensitivity 6.3 Non-linear optimization	Class notes
M7	6. Sets, Functions and Relations, Boolean Algebra 6.1 Introduction to set theory, set relations, set operators, cardinality of sets, Cartesian product of sets 6.2 Fundamentals of functions – range, domain, injection, surjection, bijection of functions 6.3 Fundamentals of relations, reflexive, symmetric and transitive properties in relations, representing relations, applications of relations, equivalence relations, partial order relations, lattices. 6.4 Boolean functions, representing Boolean functions	T2: Sec 2.1, 2.2, 2.3, 7.1 – 7.6, 10.1, 10.2
M8	7. Graph Theory 7.1 Introduction to graph theory, directed and undirected graphs, handshaking theorem, special graph structures, graph representations and isomorphism of graphs, connectedness, components, Euler, Hamilton paths and cycles	T2: Sec 8.1-8.5

Learning Outcomes:

No	Learning Outcomes
LO1	Students will be able to effectively use matrix algebra tools to analyse and solve systems of linear equations.
LO2	Students will be able to use some numerical methods to solve linear systems of equations
LO3	Students would be able to use methods in linear algebra to solve linear programming problems and methods in calculus to solve non-linear optimization problems.
LO4	Students will be able to work with some of the mathematical structures, concepts and notations used in discrete mathematics
LO5	Students will be able to apply the concepts of sets, functions, relations and graph theoretic concepts to problems in computer science

Part B: Contact Session Plan

Academic Term	I semester 2018-2019
Course Title	Mathematical Foundations for Data Science
Course No	
Lead Instructor	

Course Contents

Contact Hours	List of Topic Title	Text/Ref Book/external resource
1	Introduction to matrices, row-reduced echelon form of a matrix, Consistency of linear systems and matrix inversion • Unary and binary operations and special matrices (orthogonal matrix, upper and lower triangular, diagonal and sparse) • Row reduction and determination of rank. Comparison to computation using determinants • Use of rank in determining the consistency and inconsistency of linear systems • Row reduction to determine the inverse of the matrix (the Gauss Jordan method) (this is to be used in Simplex method later on)	T1: Sec 7.1 – 7.3, 7.5, 7.8
2	Vector space, subspace and span, Linear dependence and independence, basis and dimension, Linear transformation, rank and nullity and the rank nullity theorem • Definition and examples of vector space (R^n, space of polynomials of finite degree, n x m matrices etc.,) • Determination of whether a non-empty set of a vector space is a subspace or not • Span of a finite set	T1: Sec 7.4, 7.9 R1: Sec 3.2

	• Linear dependence and independence (theory and couple of examples) • Basis and dimension of a finite dimensional vector space • Linear transformation T: V → W (definition and a couple of examples) • Range(T) and Ker(T) as subspaces of W and V respectively • Rank Nullity Theorem (statement without proof) with examples	
3	Eigenvalues and eigenvectors of a matrix with applications • Eigenvalues – definition and method of determination of eigenvalues • Eigenvectors – definition and methods of finding the eigenvectors • Greshgorin's result on the bounds for eigenvalues • Application in conic sections with examples • Similarity transformation with examples	T1: Sec 8.2 – 8.4
4	Singular value decomposition with examples (using MATLAB) and applications (Face recognition with SVD) • SVD of a matrix (derivation) • Exemplify using matlab for a couple of matrices and also show that the singular values are arranged in descending order. • Face recognition example.	Class notes
5	Gauss elimination with scaling and partial pivoting; LU factorization and related methods • Gauss elimination (with and without scaling and partial pivoting). Take an example to shown the role played by precision. • LU factorization, Cholesky and Crout's methods with examples	T1: Sec 20.1, 20.2
6	Iterative methods of solving linear systems; Matrix eigenvalue problems and Power method for finding the dominant eigenvalue • Write Ax = b in the form (L+D+U) x = b and work out the iterative scheme for Gauss Jacobi and Gauss Seidel iterations. • Introduce vector and matrix norms (row sum, column sum and Frobenius norms) and work out a few problems in Excel / Matlab • Explain the power method and work out a couple of problems.	T1: Sec 20.3, 20.8
7 -8	Application of linear algebra in optimization. Modelling linear programming problem and the basics of Simplex algorithm and sensitivity analysis. • Model a LPP in construction of buildings. • Model the currency conversion optimization problem. • Work out the graphical method of solution in the case of 2 variable case • Simplex method for simple cases	Class notes

	• Outline how Gauss Jordan produces the inverse matrix. • Graphical sensitivity analysis (Change in objective value coefficients and rhs of constraints)	
9	Calculus of one and several variables; Limits, continuity and differentiability; Maxima and minima of functions; Steepest gradient method for finding the maximum. Constrained optimization (Lagrange multipliers) • Review limits, continuity and differentiability (graphically and algebraically) • Maxima and minima in one variable • Steepest gradient method • Lagrange multipliers (for more number of constraints)	Class notes
10	Introduction to set theory, set relations, set operators, cardinality of sets, Cartesian product of sets • Definition and examples of set and set operations • Cartesian product of sets	T2: Sec 2.1, 2.2
11	Fundamentals of functions – range, domain, injection, surjection, bijection of functions • Definition and examples of functions • Writing the range, domain and codomain for a few well known functions (exponential, logarithmic and trigonometric functions) • One-one, onto and one to one relationship definition and examples (algebraic and graphical)	T2: Sec 2.3
12	Fundamentals of relations, reflexive, symmetric and transitive properties in relations, representing relations and applications • Definition and examples of relations • Reflexive, Symmetric and Transitive properties and examples • Closures of the above relations including Warshall's algorithm	T2: Sec 7.1, 7.2, 7.3
13	Representing relations, applications of relations, equivalence relations, partial order relations, lattices. • Representation of relations as matrices • Applications of relations • Equivalence relations with examples • Partial order and well ordering • Hasse diagrams • Lattices with examples using lub and glb concepts	T2: Sec 7.4, 7.5, 7.6
14	Introduction to graph theory, directed and undirected graphs, handshaking theorem, special graph structures, graph representations • Directed and undirected graphs with examples • Handshaking theorem connecting vertices and edges • Graph representation as a matrix and paths.	T2: Sec 8.1, 8.2

15	Isomorphism of graphs, connectedness, components, Euler, Hamilton paths and cycles • Graph isomorphism definition and examples (how to show two graphs are isomorphic / non-isomorphic) • Connectedness in graphs and components • Euler path and cycle (using degree of vertices) • Hamilton path and cycle (using degree of vertices)	T2: Sec 8.3, 8.4, 8.5
16	Boolean Algebra- Boolean Functions, Representing Boolean functions and functional completeness • Definition and examples of Boolean algebras • Representation of Boolean functions • Functional completeness (Complement, union, intersection) → (complement, union) {alternatively (complement, intersection) → NAND, NOR as a single operand	T2: Sec 10.1, 10.2

 # The above contact hours and topics can be adapted for non-specific and specific WILP programs depending on the requirements and class interests.

Lab Details

Title	Access URL
Lab Setup Instructions	Not applicable
Lab Capsules	Not applicable
Additional References	Not applicable

Select Topics and Case Studies from business for experiential learning

Topic No.	Select Topics in Syllabus for experiential learning	Access URL
1	Assignment - linear algebra topics	
2	Assignment- discrete structures topics	

Evaluation Scheme

Legend: EC = Evaluation Component

No	Name	Type	Duration	Weight	Day, Date, Session, Time
1	Assignment 1	Online		10%	TBD
2	Assignment 2	Online		10%	TBD
3	Quiz 1	Online	30 min	5%	TBD
4	Quiz 2	Online	30 min	5%	TBD
5	Mid-Semester Exam	Closed book	90 min	30%	TBD
6	Comprehensive Exam	Open book	150 min	40%	TBD

Note - *Evaluation components can be tailored depending on the proposed model.*

Important Information

Syllabus for Mid-Semester Test (Closed Book): Topics in Weeks 1-8

Syllabus for Comprehensive Exam (Open Book): All topics (in sessions 1 to 16) given in plan of study

Evaluation Guidelines:

1. EC-1 consists of two Assignments and two Quizzes. Announcements regarding the same will be made in a timely manner.
2. For Closed Book tests: No books or reference material of any kind will be permitted. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
3. For Open Book exams: Use of prescribed and reference text books, in original (not photocopies) is permitted. Class notes/slides as reference material in filed or bound form is permitted. However, loose sheets of paper will not be allowed. Use of calculators is permitted in all exams. Laptops/Mobiles of any kind are not allowed. Exchange of any material is not allowed.
4. If a student is unable to appear for the Regular Test/Exam due to genuine exigencies, the student should follow the procedure to apply for the Make-Up Test/Exam. The genuineness of the reason for absence in the Regular Exam shall be assessed prior to giving permission to appear for the Make-up Exam. Make-Up Test/Exam will be conducted only at selected exam centres on the dates to be announced later.

It shall be the responsibility of the individual student to be regular in maintaining the self-study schedule as given in the course handout, attend the lectures, and take all the prescribed evaluation components such as Assignment/Quiz, Mid-Semester Test and Comprehensive Exam according to the evaluation scheme provided in the handout.

CHAPTER 7

Linear Algebra: Matrices, Vectors, Determinants. Linear Systems

Linear algebra is a fairly extensive subject that covers vectors and matrices, determinants, systems of linear equations, vector spaces and linear transformations, eigenvalue problems, and other topics. As an area of study it has a broad appeal in that it has many applications in engineering, physics, geometry, computer science, economics, and other areas. It also contributes to a deeper understanding of mathematics itself.

Matrices, which are rectangular arrays of numbers or functions, and **vectors** are the main tools of linear algebra. Matrices are important because they let us express large amounts of data and functions in an organized and concise form. Furthermore, since matrices are single objects, we denote them by single letters and calculate with them directly. All these features have made matrices and vectors very popular for expressing scientific and mathematical ideas.

The chapter keeps a good mix between applications (electric networks, Markov processes, traffic flow, etc.) and theory. Chapter 7 is structured as follows: Sections 7.1 and 7.2 provide an intuitive introduction to matrices and vectors and their operations, including matrix multiplication. The next block of sections, that is, Secs. 7.3–7.5 provide the most important method for ***solving systems of linear equations*** by the Gauss elimination method. This method is a cornerstone of linear algebra, and the method itself and variants of it appear in different areas of mathematics and in many applications. It leads to a consideration of the behavior of solutions and concepts such as rank of a matrix, linear independence, and bases. We shift to determinants, a topic that has declined in importance, in Secs. 7.6 and 7.7. Section 7.8 covers inverses of matrices. The chapter ends with vector spaces, inner product spaces, linear transformations, and composition of linear transformations. Eigenvalue problems follow in Chap. 8.

COMMENT. ***Numeric linear algebra (Secs. 20.1–20.5) can be studied immediately after this chapter.***

Prerequisite: None.
Sections that may be omitted in a short course: 7.5, 7.9.
References and Answers to Problems: App. 1 Part B, and App. 2.

www.EngineeringEBooksPdf.com

7.1 Matrices, Vectors: Addition and Scalar Multiplication

The basic concepts and rules of matrix and vector algebra are introduced in Secs. 7.1 and 7.2 and are followed by **linear systems** (systems of linear equations), a main application, in Sec. 7.3.

Let us first take a leisurely look at matrices before we formalize our discussion. A **matrix** is a rectangular array of numbers or functions which we will enclose in brackets. For example,

$$(1)\qquad \begin{bmatrix} 0.3 & 1 & -5 \\ 0 & -0.2 & 16 \end{bmatrix}, \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$
$$\begin{bmatrix} e^{-x} & 2x^2 \\ e^{6x} & 4x \end{bmatrix}, \quad [a_1 \quad a_2 \quad a_3], \quad \begin{bmatrix} 4 \\ \frac{1}{2} \end{bmatrix}$$

are matrices. The numbers (or functions) are called **entries** or, less commonly, *elements* of the matrix. The first matrix in (1) has two **rows**, which are the horizontal lines of entries. Furthermore, it has three **columns**, which are the vertical lines of entries. The second and third matrices are **square matrices**, which means that each has as many rows as columns—3 and 2, respectively. The entries of the second matrix have two indices, signifying their location within the matrix. The first index is the number of the row and the second is the number of the column, so that together the entry's position is uniquely identified. For example, a_{23} (read *a two three*) is in Row 2 and Column 3, etc. The notation is standard and applies to all matrices, including those that are not square.

Matrices having just a single row or column are called **vectors**. Thus, the fourth matrix in (1) has just one row and is called a **row vector**. The last matrix in (1) has just one column and is called a **column vector**. Because the goal of the indexing of entries was to uniquely identify the position of an element within a matrix, one index suffices for vectors, whether they are row or column vectors. Thus, the third entry of the row vector in (1) is denoted by a_3.

Matrices are handy for storing and processing data in applications. Consider the following two common examples.

EXAMPLE 1 Linear Systems, a Major Application of Matrices

We are given a system of linear equations, briefly a **linear system**, such as

$$\begin{aligned} 4x_1 + 6x_2 + 9x_3 &= 6 \\ 6x_1 \qquad\quad - 2x_3 &= 20 \\ 5x_1 - 8x_2 + x_3 &= 10 \end{aligned}$$

where x_1, x_2, x_3 are the **unknowns**. We form the **coefficient matrix**, call it **A**, by listing the coefficients of the unknowns in the position in which they appear in the linear equations. In the second equation, there is no unknown x_2, which means that the coefficient of x_2 is 0 and hence in matrix **A**, $a_{22} = 0$, Thus,

www.EngineeringEBooksPdf.com

$$\mathbf{A} = \begin{bmatrix} 4 & 6 & 9 \\ 6 & 0 & -2 \\ 5 & -8 & 1 \end{bmatrix}. \quad \text{We form another matrix} \quad \tilde{\mathbf{A}} = \begin{bmatrix} 4 & 6 & 9 & 6 \\ 6 & 0 & -2 & 20 \\ 5 & -8 & 1 & 10 \end{bmatrix}$$

by augmenting **A** with the right sides of the linear system and call it the augmented matrix of the system.

Since we can go back and recapture the system of linear equations directly from the augmented matrix $\tilde{\mathbf{A}}$, $\tilde{\mathbf{A}}$ contains all the information of the system and can thus be used to solve the linear system. This means that we can just use the augmented matrix to do the calculations needed to solve the system. We shall explain this in detail in Sec. 7.3. Meanwhile you may verify by substitution that the solution is $x_1 = 3$, $x_2 = \frac{1}{2}$, $x_3 = -1$.

The notation x_1, x_2, x_3 for the unknowns is practical but not essential; we could choose x, y, z or some other letters. ■

EXAMPLE 2 Sales Figures in Matrix Form

Sales figures for three products I, II, III in a store on Monday (Mon), Tuesday (Tues), $\cdots$ may for each week be arranged in a matrix

$$\begin{array}{cc} & \begin{array}{ccccccc} \text{Mon} & \text{Tues} & \text{Wed} & \text{Thur} & \text{Fri} & \text{Sat} & \text{Sun} \end{array} \\ \mathbf{A} = & \begin{bmatrix} 40 & 33 & 81 & 0 & 21 & 47 & 33 \\ 0 & 12 & 78 & 50 & 50 & 96 & 90 \\ 10 & 0 & 0 & 27 & 43 & 78 & 56 \end{bmatrix} \begin{array}{c} \text{I} \\ \cdot\ \text{II} \\ \text{III} \end{array} \end{array}$$

If the company has 10 stores, we can set up 10 such matrices, one for each store. Then, by adding corresponding entries of these matrices, we can get a matrix showing the total sales of each product on each day. Can you think of other data which can be stored in matrix form? For instance, in transportation or storage problems? Or in listing distances in a network of roads? ■

General Concepts and Notations

Let us formalize what we just have discussed. We shall denote matrices by capital boldface letters **A**, **B**, **C**, $\cdots$, or by writing the general entry in brackets; thus $\mathbf{A} = [a_{jk}]$, and so on. By an ***m* × *n* matrix** (read *m by n matrix*) we mean a matrix with m rows and n columns—rows always come first! $m \times n$ is called the **size** of the matrix. Thus an $m \times n$ matrix is of the form

$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}. \tag{2}$$

The matrices in (1) are of sizes 2×3, 3×3, 2×2, 1×3, and 2×1, respectively.

Each entry in (2) has two subscripts. The first is the *row number* and the second is the *column number*. Thus a_{21} is the entry in Row 2 and Column 1.

If $m = n$, we call **A** an $n \times n$ **square matrix**. Then its diagonal containing the entries $a_{11}, a_{22}, \cdots, a_{nn}$ is called the **main diagonal** of **A**. Thus the main diagonals of the two square matrices in (1) are a_{11}, a_{22}, a_{33} and e^{-x}, $4x$, respectively.

Square matrices are particularly important, as we shall see. A matrix of any size $m \times n$ is called a **rectangular matrix**; this includes square matrices as a special case.

www.EngineeringEBooksPdf.com

Vectors

A **vector** is a matrix with only one row or column. Its entries are called the **components** of the vector. We shall denote vectors by ***lowercase*** boldface letters $\mathbf{a}$, $\mathbf{b}, \cdots$ or by its general component in brackets, $\mathbf{a} = [a_j]$, and so on. Our special vectors in (1) suggest that a (general) **row vector** is of the form

$$\mathbf{a} = [a_1 \quad a_2 \quad \cdots \quad a_n]. \qquad \text{For instance,} \qquad \mathbf{a} = [-2 \quad 5 \quad 0.8 \quad 0 \quad 1].$$

A **column vector** is of the form

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}. \qquad \text{For instance,} \qquad \mathbf{b} = \begin{bmatrix} 4 \\ 0 \\ -7 \end{bmatrix}.$$

Addition and Scalar Multiplication of Matrices and Vectors

What makes matrices and vectors really useful and particularly suitable for computers is the fact that we can calculate with them almost as easily as with numbers. Indeed, we now introduce rules for addition and for scalar multiplication (multiplication by numbers) that were suggested by practical applications. (Multiplication of matrices by matrices follows in the next section.) We first need the concept of equality.

DEFINITION

Equality of Matrices

Two matrices $\mathbf{A} = [a_{jk}]$ and $\mathbf{B} = [b_{jk}]$ are **equal**, written $\mathbf{A} = \mathbf{B}$, if and only if they have the same size and the corresponding entries are equal, that is, $a_{11} = b_{11}$, $a_{12} = b_{12}$, and so on. Matrices that are not equal are called **different**. Thus, matrices of different sizes are always different.

EXAMPLE 3 Equality of Matrices

Let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \text{and} \qquad \mathbf{B} = \begin{bmatrix} 4 & 0 \\ 3 & -1 \end{bmatrix}.$$

Then

$$\mathbf{A} = \mathbf{B} \qquad \text{if and only if} \qquad \begin{aligned} a_{11} &= 4, \quad a_{12} = 0, \\ a_{21} &= 3, \quad a_{22} = -1. \end{aligned}$$

The following matrices are all different. Explain!

$$\begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix} \qquad \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} \qquad \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \qquad \begin{bmatrix} 1 & 3 & 0 \\ 4 & 2 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 & 3 \\ 0 & 4 & 2 \end{bmatrix}$$

■

www.EngineeringEBooksPdf.com

DEFINITION

Addition of Matrices

The **sum** of two matrices $\mathbf{A} = [a_{jk}]$ and $\mathbf{B} = [b_{jk}]$ ***of the same size*** is written $\mathbf{A} + \mathbf{B}$ and has the entries $a_{jk} + b_{jk}$ obtained by adding the corresponding entries of $\mathbf{A}$ and $\mathbf{B}$. Matrices of different sizes cannot be added.

As a special case, the **sum** $\mathbf{a} + \mathbf{b}$ of two row vectors or two column vectors, which must have the same number of components, is obtained by adding the corresponding components.

EXAMPLE 4 Addition of Matrices and Vectors

$$\text{If} \quad \mathbf{A} = \begin{bmatrix} -4 & 6 & 3 \\ 0 & 1 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 5 & -1 & 0 \\ 3 & 1 & 0 \end{bmatrix}, \quad \text{then} \quad \mathbf{A} + \mathbf{B} = \begin{bmatrix} 1 & 5 & 3 \\ 3 & 2 & 2 \end{bmatrix}.$$

$\mathbf{A}$ in Example 3 and our present $\mathbf{A}$ cannot be added. If $\mathbf{a} = [5 \quad 7 \quad 2]$ and $\mathbf{b} = [-6 \quad 2 \quad 0]$, then $\mathbf{a} + \mathbf{b} = [-1 \quad 9 \quad 2]$.

An application of matrix addition was suggested in Example 2. Many others will follow. ■

DEFINITION

Scalar Multiplication (Multiplication by a Number)

The **product** of any $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ and any **scalar** c (number c) is written $c\mathbf{A}$ and is the $m \times n$ matrix $c\mathbf{A} = [ca_{jk}]$ obtained by multiplying each entry of $\mathbf{A}$ by c.

Here $(-1)\mathbf{A}$ is simply written $-\mathbf{A}$ and is called the **negative** of $\mathbf{A}$. Similarly, $(-k)\mathbf{A}$ is written $-k\mathbf{A}$. Also, $\mathbf{A} + (-\mathbf{B})$ is written $\mathbf{A} - \mathbf{B}$ and is called the **difference** of $\mathbf{A}$ and $\mathbf{B}$ (which must have the same size!).

EXAMPLE 5 Scalar Multiplication

$$\text{If} \quad \mathbf{A} = \begin{bmatrix} 2.7 & -1.8 \\ 0 & 0.9 \\ 9.0 & -4.5 \end{bmatrix}, \quad \text{then} \quad -\mathbf{A} = \begin{bmatrix} -2.7 & 1.8 \\ 0 & -0.9 \\ -9.0 & 4.5 \end{bmatrix}, \quad \frac{10}{9}\mathbf{A} = \begin{bmatrix} 3 & -2 \\ 0 & 1 \\ 10 & -5 \end{bmatrix}, \quad 0\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

If a matrix $\mathbf{B}$ shows the distances between some cities in miles, $1.609\mathbf{B}$ gives these distances in kilometers. ■

Rules for Matrix Addition and Scalar Multiplication. From the familiar laws for the addition of numbers we obtain similar laws for the addition of matrices of the same size $m \times n$, namely,

$$\textbf{(3)} \qquad \begin{aligned} &\text{(a)} \quad \mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \\ &\text{(b)} \quad (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \qquad \text{(written } \mathbf{A} + \mathbf{B} + \mathbf{C}\text{)} \\ &\text{(c)} \quad \mathbf{A} + \mathbf{0} = \mathbf{A} \\ &\text{(d)} \quad \mathbf{A} + (-\mathbf{A}) = \mathbf{0}. \end{aligned}$$

Here $\mathbf{0}$ denotes the **zero matrix** (of size $m \times n$), that is, the $m \times n$ matrix with all entries zero. If $m = 1$ or $n = 1$, this is a vector, called a **zero vector**.

www.EngineeringEBooksPdf.com

Hence matrix addition is *commutative* and *associative* [by (3a) and (3b)]. Similarly, for scalar multiplication we obtain the rules

$$\textbf{(4)} \qquad \begin{aligned} &\text{(a)} \quad c(\mathbf{A}+\mathbf{B}) = c\mathbf{A} + c\mathbf{B} \\ &\text{(b)} \quad (c+k)\mathbf{A} = c\mathbf{A} + k\mathbf{A} \\ &\text{(c)} \quad c(k\mathbf{A}) = (ck)\mathbf{A} \qquad \text{(written } ck\mathbf{A}\text{)} \\ &\text{(d)} \quad 1\mathbf{A} = \mathbf{A}. \end{aligned}$$

PROBLEM SET 7.1

1–7 GENERAL QUESTIONS

1. **Equality.** Give reasons why the five matrices in Example 3 are all different.
2. **Double subscript notation.** If you write the matrix in Example 2 in the form $\mathbf{A} = [a_{jk}]$, what is a_{31}? a_{13}? a_{26}? a_{33}?
3. **Sizes.** What sizes do the matrices in Examples 1, 2, 3, and 5 have?
4. **Main diagonal.** What is the main diagonal of **A** in Example 1? Of **A** and **B** in Example 3?
5. **Scalar multiplication.** If **A** in Example 2 shows the number of items sold, what is the matrix **B** of units sold if a unit consists of **(a)** 5 items and **(b)** 10 items?
6. If a 12×12 matrix **A** shows the distances between 12 cities in kilometers, how can you obtain from **A** the matrix **B** showing these distances in miles?
7. **Addition of vectors.** Can you add: A row and a column vector with different numbers of components? With the same number of components? Two row vectors with the same number of components but different numbers of zeros? A vector and a scalar? A vector with four components and a 2×2 matrix?

8–16 ADDITION AND SCALAR MULTIPLICATION OF MATRICES AND VECTORS

Let

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & 4 \\ 6 & 5 & 5 \\ 1 & 0 & -3 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 5 & 2 \\ 5 & 3 & 4 \\ -2 & 4 & -2 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 5 & 2 \\ -2 & 4 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} -4 & 1 \\ 5 & 0 \\ 2 & -1 \end{bmatrix},$$

$$\mathbf{E} = \begin{bmatrix} 0 & 2 \\ 3 & 4 \\ 3 & -1 \end{bmatrix}$$

$$\mathbf{u} = \begin{bmatrix} 1.5 \\ 0 \\ -3.0 \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} -1 \\ 3 \\ 2 \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} -5 \\ -30 \\ 10 \end{bmatrix}.$$

Find the following expressions, indicating which of the rules in (3) or (4) they illustrate, or give reasons why they are not defined.

8. $2\mathbf{A} + 4\mathbf{B}, \quad 4\mathbf{B} + 2\mathbf{A}, \quad 0\mathbf{A} + \mathbf{B}, \quad 0.4\mathbf{B} - 4.2\mathbf{A}$
9. $3\mathbf{A}, \quad 0.5\mathbf{B}, \quad 3\mathbf{A} + 0.5\mathbf{B}, \quad 3\mathbf{A} + 0.5\mathbf{B} + \mathbf{C}$
10. $(4 \cdot 3)\mathbf{A}, \quad 4(3\mathbf{A}), \quad 14\mathbf{B} - 3\mathbf{B}, \quad 11\mathbf{B}$
11. $8\mathbf{C} + 10\mathbf{D}, \quad 2(5\mathbf{D} + 4\mathbf{C}), \quad 0.6\mathbf{C} - 0.6\mathbf{D}, \quad 0.6(\mathbf{C} - \mathbf{D})$
12. $(\mathbf{C} + \mathbf{D}) + \mathbf{E}, \quad (\mathbf{D} + \mathbf{E}) + \mathbf{C}, \quad 0(\mathbf{C} - \mathbf{E}) + 4\mathbf{D}, \quad \mathbf{A} - 0\mathbf{C}$
13. $(2 \cdot 7)\mathbf{C}, \quad 2(7\mathbf{C}), \quad -\mathbf{D} + 0\mathbf{E}, \quad \mathbf{E} - \mathbf{D} + \mathbf{C} + \mathbf{u}$
14. $(5\mathbf{u} + 5\mathbf{v}) - \frac{1}{2}\mathbf{w}, \quad -20(\mathbf{u} + \mathbf{v}) + 2\mathbf{w}, \quad \mathbf{E} - (\mathbf{u} + \mathbf{v}), \quad 10(\mathbf{u} + \mathbf{v}) + \mathbf{w}$
15. $(\mathbf{u} + \mathbf{v}) - \mathbf{w}, \quad \mathbf{u} + (\mathbf{v} - \mathbf{w}), \quad \mathbf{C} + 0\mathbf{w}, \quad 0\mathbf{E} + \mathbf{u} - \mathbf{v}$
16. $15\mathbf{v} - 3\mathbf{w} - 0\mathbf{u}, \quad -3\mathbf{w} + 15\mathbf{v}, \quad \mathbf{D} - \mathbf{u} + 3\mathbf{C}, \quad 8.5\mathbf{w} - 11.1\mathbf{u} + 0.4\mathbf{v}$
17. **Resultant of forces.** If the above vectors **u**, **v**, **w** represent forces in space, their sum is called their *resultant*. Calculate it.
18. **Equilibrium.** By definition, forces are *in equilibrium* if their resultant is the zero vector. Find a force **p** such that the above **u**, **v**, **w**, and **p** are in equilibrium.
19. **General rules.** Prove (3) and (4) for general 2×3 matrices and scalars c and k.

www.EngineeringEBooksPdf.com

20. TEAM PROJECT. Matrices for Networks. Matrices have various engineering applications, as we shall see. For instance, they can be used to characterize connections in electrical networks, in nets of roads, in production processes, etc., as follows.

(a) Nodal Incidence Matrix. The network in Fig. 155 consists of six *branches* (connections) and four *nodes* (points where two or more branches come together). One node is the *reference node* (grounded node, whose voltage is zero). We number the other nodes and number and direct the branches. This we do arbitrarily. The network can now be described by a matrix $\mathbf{A} = [a_{jk}]$, where

$$a_{jk} = \begin{cases} +1 \text{ if branch } k \text{ leaves node } (j) \\ -1 \text{ if branch } k \text{ enters node } (j) \\ 0 \text{ if branch } k \text{ does not touch node } (j). \end{cases}$$

A is called the *nodal incidence matrix* of the network. Show that for the network in Fig. 155 the matrix **A** has the given form.

$$\begin{array}{lc} \text{Branch} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{l} \text{Node } (1) \\ \text{Node } (2) \\ \text{Node } (3) \end{array} & \begin{bmatrix} 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & -1 & -1 \end{bmatrix} \end{array}$$

Fig. 155. Network and nodal incidence matrix in Team Project 20(a)

(b) Find the nodal incidence matrices of the networks in Fig. 156.

Fig. 156. Electrical networks in Team Project 20(b)

(c) Sketch the three networks corresponding to the nodal incidence matrices

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & -1 & 0 & 0 & 1 \\ -1 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix},$$

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

(d) Mesh Incidence Matrix. A network can also be characterized by the *mesh incidence matrix* $\mathbf{M} = [m_{jk}]$, where

$$m_{jk} = \begin{cases} +1 \text{ if branch } k \text{ is in mesh } \boxed{j} \\ \quad \text{and has the same orientation} \\ -1 \text{ if branch } k \text{ is in mesh } \boxed{j} \\ \quad \text{and has the opposite orientation} \\ 0 \text{ if branch } k \text{ is not in mesh } \boxed{j} \end{cases}$$

and a mesh is a loop with no branch in its interior (or in its exterior). Here, the meshes are numbered and directed (oriented) in an arbitrary fashion. Show that for the network in Fig. 157, the matrix **M** has the given form, where Row 1 corresponds to mesh 1, etc.

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

Fig. 157. Network and matrix **M** in Team Project 20(d)

www.EngineeringEBooksPdf.com

7.2 Matrix Multiplication

Matrix multiplication means that one multiplies matrices by matrices. Its definition is standard but it looks artificial. *Thus you have to study matrix multiplication carefully*, multiply a few matrices together for practice until you can understand how to do it. Here then is the definition. (Motivation follows later.)

DEFINITION

Multiplication of a Matrix by a Matrix

The **product** $\mathbf{C} = \mathbf{AB}$ (in this order) of an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ times an $r \times p$ matrix $\mathbf{B} = [b_{jk}]$ is defined if and only if $r = n$ and is then the $m \times p$ matrix $\mathbf{C} = [c_{jk}]$ with entries

$$c_{jk} = \sum_{l=1}^{n} a_{jl}b_{lk} = a_{j1}b_{1k} + a_{j2}b_{2k} + \cdots + a_{jn}b_{nk} \qquad \begin{array}{l} j = 1, \cdots, m \\ k = 1, \cdots, p. \end{array} \tag{1}$$

The condition $r = n$ means that the second factor, $\mathbf{B}$, must have as many rows as the first factor has columns, namely n. A diagram of sizes that shows when matrix multiplication is possible is as follows:

$$\begin{array}{ccccc} \mathbf{A} & \mathbf{B} & = & \mathbf{C} \\ [m \times n] & [n \times p] & = & [m \times p]. \end{array}$$

The entry c_{jk} in (1) is obtained by multiplying each entry in the jth row of $\mathbf{A}$ by the corresponding entry in the kth column of $\mathbf{B}$ and then adding these n products. For instance, $c_{21} = a_{21}b_{11} + a_{22}b_{21} + \cdots + a_{2n}b_{n1}$, and so on. One calls this briefly a ***multiplication of rows into columns***. For $n = 3$, this is illustrated by

$$m = 4 \left\{ \overbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix}}^{n=3} \right. \overbrace{\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}}^{p=2} = \overbrace{\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \\ c_{41} & c_{42} \end{bmatrix}}^{p=2} \left. \right\} m = 4$$

Notations in a product $\mathbf{AB} = \mathbf{C}$

where we shaded the entries that contribute to the calculation of entry c_{21} just discussed.

Matrix multiplication will be motivated by its use in *linear transformations* in this section and more fully in Sec. 7.9.

Let us illustrate the main points of matrix multiplication by some examples. Note that matrix multiplication also includes multiplying a matrix by a vector, since, after all, a vector is a special matrix.

EXAMPLE 1 Matrix Multiplication

$$\mathbf{AB} = \begin{bmatrix} 3 & 5 & -1 \\ 4 & 0 & 2 \\ -6 & -3 & 2 \end{bmatrix} \begin{bmatrix} 2 & -2 & 3 & 1 \\ 5 & 0 & 7 & 8 \\ 9 & -4 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 22 & -2 & 43 & 42 \\ 26 & -16 & 14 & 6 \\ -9 & 4 & -37 & -28 \end{bmatrix}$$

Here $c_{11} = 3 \cdot 2 + 5 \cdot 5 + (-1) \cdot 9 = 22$, and so on. The entry in the box is $c_{23} = 4 \cdot 3 + 0 \cdot 7 + 2 \cdot 1 = 14$. The product $\mathbf{BA}$ is not defined. ■

www.EngineeringEBooksPdf.com

EXAMPLE 2 Multiplication of a Matrix and a Vector

$$\begin{bmatrix} 4 & 2 \\ 1 & 8 \end{bmatrix}\begin{bmatrix} 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 4 \cdot 3 + 2 \cdot 5 \\ 1 \cdot 3 + 8 \cdot 5 \end{bmatrix} = \begin{bmatrix} 22 \\ 43 \end{bmatrix} \quad \text{whereas} \quad \begin{bmatrix} 3 \\ 5 \end{bmatrix}\begin{bmatrix} 4 & 2 \\ 1 & 8 \end{bmatrix} \quad \text{is undefined.}$$

EXAMPLE 3 Products of Row and Column Vectors

$$\begin{bmatrix} 3 & 6 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} = [19], \qquad \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}\begin{bmatrix} 3 & 6 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 1 \\ 6 & 12 & 2 \\ 12 & 24 & 4 \end{bmatrix}.$$

EXAMPLE 4 CAUTION! Matrix Multiplication Is Not Commutative, AB ≠ BA in General

This is illustrated by Examples 1 and 2, where one of the two products is not even defined, and by Example 3, where the two products have different sizes. But it also holds for square matrices. For instance,

$$\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{but} \quad \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix} = \begin{bmatrix} 99 & 99 \\ -99 & -99 \end{bmatrix}.$$

It is interesting that this also shows that $\mathbf{AB} = \mathbf{0}$ does ***not*** necessarily imply $\mathbf{BA} = \mathbf{0}$ or $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$. We shall discuss this further in Sec. 7.8, along with reasons when this happens.

Our examples show that in matrix products ***the order of factors must always be observed very carefully***. Otherwise matrix multiplication satisfies rules similar to those for numbers, namely.

$$(2) \qquad \begin{aligned} &\text{(a)} \quad (k\mathbf{A})\mathbf{B} = k(\mathbf{AB}) = \mathbf{A}(k\mathbf{B}) \quad \textit{written } k\mathbf{AB} \textit{ or } \mathbf{A}k\mathbf{B} \\ &\text{(b)} \quad \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C} \quad \textit{written } \mathbf{ABC} \\ &\text{(c)} \quad (\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC} \\ &\text{(d)} \quad \mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB} \end{aligned}$$

provided **A**, **B**, and **C** are such that the expressions on the left are defined; here, k is any scalar. (2b) is called the **associative law**. (2c) and (2d) are called the **distributive laws**.

Since matrix multiplication is a multiplication of rows into columns, we can write the defining formula (1) more compactly as

$$\textbf{(3)} \qquad c_{jk} = \mathbf{a}_j\mathbf{b}_k, \qquad j = 1, \cdots, m; \quad k = 1, \cdots, p,$$

where $\mathbf{a}_j$ is the jth row vector of **A** and $\mathbf{b}_k$ is the kth column vector of **B**, so that in agreement with (1),

$$\mathbf{a}_j\mathbf{b}_k = \begin{bmatrix} a_{j1} & a_{j2} & \cdots & a_{jn} \end{bmatrix}\begin{bmatrix} b_{1k} \\ \vdots \\ b_{nk} \end{bmatrix} = a_{j1}b_{1k} + a_{j2}b_{2k} + \cdots + a_{jn}b_{nk}.$$

www.EngineeringEBooksPdf.com

EXAMPLE 5 Product in Terms of Row and Column Vectors

If $\mathbf{A} = [a_{jk}]$ is of size 3×3 and $\mathbf{B} = [b_{jk}]$ is of size 3×4, then

$$\mathbf{AB} = \begin{bmatrix} \mathbf{a}_1\mathbf{b}_1 & \mathbf{a}_1\mathbf{b}_2 & \mathbf{a}_1\mathbf{b}_3 & \mathbf{a}_1\mathbf{b}_4 \\ \mathbf{a}_2\mathbf{b}_1 & \mathbf{a}_2\mathbf{b}_2 & \mathbf{a}_2\mathbf{b}_3 & \mathbf{a}_2\mathbf{b}_4 \\ \mathbf{a}_3\mathbf{b}_1 & \mathbf{a}_3\mathbf{b}_2 & \mathbf{a}_3\mathbf{b}_3 & \mathbf{a}_3\mathbf{b}_4 \end{bmatrix}. \tag{4}$$

Taking $\mathbf{a}_1 = [3 \quad 5 \quad -1]$, $\mathbf{a}_2 = [4 \quad 0 \quad 2]$, etc., verify (4) for the product in Example 1. ■

Parallel processing of products on the computer is facilitated by a variant of (3) for computing $\mathbf{C} = \mathbf{AB}$, which is used by standard algorithms (such as in Lapack). In this method, $\mathbf{A}$ is used as given, $\mathbf{B}$ is taken in terms of its column vectors, and the product is computed columnwise; thus,

$$\mathbf{AB} = \mathbf{A}[\mathbf{b}_1 \quad \mathbf{b}_2 \quad \cdots \quad \mathbf{b}_p] = [\mathbf{Ab}_1 \quad \mathbf{Ab}_2 \quad \cdots \quad \mathbf{Ab}_p]. \tag{5}$$

Columns of $\mathbf{B}$ are then assigned to different processors (individually or several to each processor), which simultaneously compute the columns of the product matrix $\mathbf{Ab}_1$, $\mathbf{Ab}_2$, etc.

EXAMPLE 6 Computing Products Columnwise by (5)

To obtain

$$\mathbf{AB} = \begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 3 & 0 & 7 \\ -1 & 4 & 6 \end{bmatrix} = \begin{bmatrix} 11 & 4 & 34 \\ -17 & 8 & -23 \end{bmatrix}$$

from (5), calculate the columns

$$\begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} 11 \\ -17 \end{bmatrix}, \quad \begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \end{bmatrix}, \quad \begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 7 \\ 6 \end{bmatrix} = \begin{bmatrix} 34 \\ -23 \end{bmatrix}$$

of $\mathbf{AB}$ and then write them as a single matrix, as shown in the first formula on the right. ■

Motivation of Multiplication by Linear Transformations

Let us now motivate the "unnatural" matrix multiplication by its use in **linear transformations**. For $n = 2$ variables these transformations are of the form

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 \\ y_2 &= a_{21}x_1 + a_{22}x_2 \end{aligned} \tag{6*}$$

and suffice to explain the idea. (For general n they will be discussed in Sec. 7.9.) For instance, (6*) may relate an x_1x_2-coordinate system to a y_1y_2-coordinate system in the plane. In vectorial form we can write (6*) as

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{Ax} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix}. \tag{6}$$

www.EngineeringEBooksPdf.com

Now suppose further that the x_1x_2-system is related to a w_1w_2-system by another linear transformation, say,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{Bw} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} b_{11}w_1 + b_{12}w_2 \\ b_{21}w_1 + b_{22}w_2 \end{bmatrix}. \tag{7}$$

Then the y_1y_2-system is related to the w_1w_2-system indirectly via the x_1x_2-system, and we wish to express this relation directly. Substitution will show that this direct relation is a linear transformation, too, say,

$$\mathbf{y} = \mathbf{Cw} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} c_{11}w_1 + c_{12}w_2 \\ c_{21}w_1 + c_{22}w_2 \end{bmatrix}. \tag{8}$$

Indeed, substituting (7) into (6), we obtain

$$\begin{aligned} y_1 &= a_{11}(b_{11}w_1 + b_{12}w_2) + a_{12}(b_{21}w_1 + b_{22}w_2) \\ &= (a_{11}b_{11} + a_{12}b_{21})w_1 + (a_{11}b_{12} + a_{12}b_{22})w_2 \\ y_2 &= a_{21}(b_{11}w_1 + b_{12}w_2) + a_{22}(b_{21}w_1 + b_{22}w_2) \\ &= (a_{21}b_{11} + a_{22}b_{21})w_1 + (a_{21}b_{12} + a_{22}b_{22})w_2. \end{aligned}$$

Comparing this with (8), we see that

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21} \qquad & c_{12} &= a_{11}b_{12} + a_{12}b_{22} \\ c_{21} &= a_{21}b_{11} + a_{22}b_{21} \qquad & c_{22} &= a_{21}b_{12} + a_{22}b_{22}. \end{aligned}$$

This proves that $\mathbf{C} = \mathbf{AB}$ with the product defined as in (1). For larger matrix sizes the idea and result are exactly the same. Only the number of variables changes. We then have m variables y and n variables x and p variables w. The matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C} = \mathbf{AB}$ then have sizes $m \times n$, $n \times p$, and $m \times p$, respectively. And the requirement that $\mathbf{C}$ be the product $\mathbf{AB}$ leads to formula (1) in its general form. *This motivates matrix multiplication.*

Transposition

We obtain the transpose of a matrix by writing its rows as columns (or equivalently its columns as rows). This also applies to the transpose of vectors. Thus, a row vector becomes a column vector and vice versa. In addition, for square matrices, we can also "reflect" the elements along the main diagonal, that is, interchange entries that are symmetrically positioned with respect to the main diagonal to obtain the transpose. Hence a_{12} becomes a_{21}, a_{31} becomes a_{13}, and so forth. Example 7 illustrates these ideas. Also note that, if $\mathbf{A}$ is the given matrix, then we denote its transpose by $\mathbf{A}^\mathsf{T}$.

EXAMPLE 7 Transposition of Matrices and Vectors

If
$$\mathbf{A} = \begin{bmatrix} 5 & -8 & 1 \\ 4 & 0 & 0 \end{bmatrix}, \quad \text{then} \quad \mathbf{A}^\mathsf{T} = \begin{bmatrix} 5 & 4 \\ -8 & 0 \\ 1 & 0 \end{bmatrix}.$$

www.EngineeringEBooksPdf.com

A little more compactly, we can write

$$\begin{bmatrix} 5 & -8 & 1 \\ 4 & 0 & 0 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} 5 & 4 \\ -8 & 0 \\ 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 3 & 0 & 7 \\ 8 & -1 & 5 \\ 1 & -9 & 4 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} 3 & 8 & 1 \\ 0 & -1 & -9 \\ 7 & 5 & 4 \end{bmatrix},$$

Furthermore, the transpose $[6 \quad 2 \quad 3]^{\mathsf{T}}$ of the row vector $[6 \quad 2 \quad 3]$ is the column vector

$$[6 \quad 2 \quad 3]^{\mathsf{T}} = \begin{bmatrix} 6 \\ 2 \\ 3 \end{bmatrix}. \qquad \text{Conversely,} \qquad \begin{bmatrix} 6 \\ 2 \\ 3 \end{bmatrix}^{\mathsf{T}} = [6 \quad 2 \quad 3].$$

■

DEFINITION

Transposition of Matrices and Vectors

The transpose of an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the $n \times m$ matrix $\mathbf{A}^{\mathsf{T}}$ (read *A transpose*) that has the first *row* of $\mathbf{A}$ as its first *column*, the second *row* of $\mathbf{A}$ as its second *column*, and so on. Thus the transpose of $\mathbf{A}$ in (2) is $\mathbf{A}^{\mathsf{T}} = [a_{kj}]$, written out

$$\mathbf{A}^{\mathsf{T}} = [a_{kj}] = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdot & \cdot & \cdots & \cdot \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}. \tag{9}$$

As a special case, transposition converts row vectors to column vectors and conversely.

Transposition gives us a choice in that we can work either with the matrix or its transpose, whichever is more convenient.

Rules for transposition are

$$\begin{aligned} &\text{(a)} \quad (\mathbf{A}^{\mathsf{T}})^{\mathsf{T}} = \mathbf{A} \\ &\text{(b)} \quad (\mathbf{A} + \mathbf{B})^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} + \mathbf{B}^{\mathsf{T}} \\ &\text{(c)} \quad (c\mathbf{A})^{\mathsf{T}} = c\mathbf{A}^{\mathsf{T}} \\ &\text{(d)} \quad (\mathbf{AB})^{\mathsf{T}} = \mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}. \end{aligned} \tag{10}$$

CAUTION! Note that in (10d) the transposed matrices are ***in reversed order.*** We leave the proofs as an exercise in Probs. 9 and 10.

Special Matrices

Certain kinds of matrices will occur quite frequently in our work, and we now list the most important ones of them.

Symmetric and Skew-Symmetric Matrices. Transposition gives rise to two useful classes of matrices. **Symmetric** matrices are square matrices whose transpose equals the

www.EngineeringEBooksPdf.com

matrix itself. **Skew-symmetric** matrices are square matrices whose transpose equals ***minus*** the matrix. Both cases are defined in (11) and illustrated by Example 8.

$$\textbf{(11)} \qquad \underbrace{\mathbf{A}^\mathsf{T} = \mathbf{A} \quad (\text{thus } a_{kj} = a_{jk}),}_{\text{Symmetric Matrix}} \qquad \underbrace{\mathbf{A}^\mathsf{T} = -\mathbf{A} \quad (\text{thus } a_{kj} = -a_{jk}, \text{ hence } a_{jj} = 0).}_{\text{Skew-Symmetric Matrix}}$$

EXAMPLE 8 Symmetric and Skew-Symmetric Matrices

$$\mathbf{A} = \begin{bmatrix} 20 & 120 & 200 \\ 120 & 10 & 150 \\ 200 & 150 & 30 \end{bmatrix} \quad \text{is symmetric, and} \quad \mathbf{B} = \begin{bmatrix} 0 & 1 & -3 \\ -1 & 0 & -2 \\ 3 & 2 & 0 \end{bmatrix} \quad \text{is skew-symmetric.}$$

For instance, if a company has three building supply centers C_1, C_2, C_3, then $\mathbf{A}$ could show costs, say, a_{jj} for handling 1000 bags of cement at center C_j, and a_{jk} $(j \neq k)$ the cost of shipping 1000 bags from C_j to C_k. Clearly, $a_{jk} = a_{kj}$ if we assume shipping in the opposite direction will cost the same.

Symmetric matrices have several general properties which make them important. This will be seen as we proceed. ■

Triangular Matrices. **Upper triangular matrices** are square matrices that can have nonzero entries only on and ***above*** the main diagonal, whereas any entry below the diagonal must be zero. Similarly, **lower triangular matrices** can have nonzero entries only on and ***below*** the main diagonal. Any entry on the main diagonal of a triangular matrix may be zero or not.

EXAMPLE 9 Upper and Lower Triangular Matrices

$$\underbrace{\begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}, \quad \begin{bmatrix} 1 & 4 & 2 \\ 0 & 3 & 2 \\ 0 & 0 & 6 \end{bmatrix}}_{\text{Upper triangular}}, \qquad \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 8 & -1 & 0 \\ 7 & 6 & 8 \end{bmatrix}, \quad \begin{bmatrix} 3 & 0 & 0 & 0 \\ 9 & -3 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 9 & 3 & 6 \end{bmatrix}}_{\text{Lower triangular}}. \quad ■$$

Diagonal Matrices. These are square matrices that can have nonzero entries only on the main diagonal. Any entry above or below the main diagonal must be zero.

If all the diagonal entries of a diagonal matrix $\mathbf{S}$ are equal, say, c, we call $\mathbf{S}$ a **scalar matrix** because multiplication of any square matrix $\mathbf{A}$ of the same size by $\mathbf{S}$ has the same effect as the multiplication by a scalar, that is,

$$\mathbf{AS} = \mathbf{SA} = c\mathbf{A}. \tag{12}$$

In particular, a scalar matrix, whose entries on the main diagonal are all 1, is called a **unit matrix** (or **identity matrix**) and is denoted by $\mathbf{I}_n$ or simply by $\mathbf{I}$. For $\mathbf{I}$, formula (12) becomes

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}. \tag{13}$$

EXAMPLE 10 Diagonal Matrix D. Scalar Matrix S. Unit Matrix I

$$\mathbf{D} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} c & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & c \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad ■$$

www.EngineeringEBooksPdf.com

Some Applications of Matrix Multiplication

EXAMPLE 11 Computer Production. Matrix Times Matrix

Supercomp Ltd produces two computer models PC1086 and PC1186. The matrix **A** shows the cost per computer (in thousands of dollars) and **B** the production figures for the year 2010 (in multiples of 10,000 units.) Find a matrix **C** that shows the shareholders the cost per quarter (in millions of dollars) for raw material, labor, and miscellaneous.

$$\mathbf{A} = \begin{array}{c} \begin{array}{cc} \text{PC1086} & \text{PC1186} \end{array} \\ \begin{bmatrix} 1.2 & 1.6 \\ 0.3 & 0.4 \\ 0.5 & 0.6 \end{bmatrix} \end{array} \begin{array}{l} \\ \text{Raw Components} \\ \text{Labor} \\ \text{Miscellaneous} \end{array} \qquad \mathbf{B} = \begin{array}{c} \text{Quarter} \\ \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 3 & 8 & 6 & 9 \\ 6 & 2 & 4 & 3 \end{bmatrix} \end{array} \begin{array}{l} \\ \text{PC1086} \\ \text{PC1186} \end{array}$$

Solution.

$$\mathbf{C} = \mathbf{AB} = \begin{array}{c} \text{Quarter} \\ \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 13.2 & 12.8 & 13.6 & 15.6 \\ 3.3 & 3.2 & 3.4 & 3.9 \\ 5.1 & 5.2 & 5.4 & 6.3 \end{bmatrix} \end{array} \begin{array}{l} \\ \text{Raw Components} \\ \text{Labor} \\ \text{Miscellaneous} \end{array}$$

Since cost is given in multiples of \$1000 and production in multiples of 10,000 units, the entries of **C** are multiples of \$10 millions; thus $c_{11} = 13.2$ means \$132 million, etc. ■

EXAMPLE 12 Weight Watching. Matrix Times Vector

Suppose that in a weight-watching program, a person of 185 lb burns 350 cal/hr in walking (3 mph), 500 in bicycling (13 mph), and 950 in jogging (5.5 mph). Bill, weighing 185 lb, plans to exercise according to the matrix shown. Verify the calculations (W = Walking, B = Bicycling, J = Jogging).

$$\begin{array}{c} \\ \text{MON} \\ \text{WED} \\ \text{FRI} \\ \text{SAT} \end{array} \begin{array}{c} \begin{array}{ccc} \text{W} & \text{B} & \text{J} \end{array} \\ \begin{bmatrix} 1.0 & 0 & 0.5 \\ 1.0 & 1.0 & 0.5 \\ 1.5 & 0 & 0.5 \\ 2.0 & 1.5 & 1.0 \end{bmatrix} \end{array} \begin{bmatrix} 350 \\ 500 \\ 950 \end{bmatrix} = \begin{bmatrix} 825 \\ 1325 \\ 1000 \\ 2400 \end{bmatrix} \begin{array}{l} \\ \text{MON} \\ \text{WED} \\ \text{FRI} \\ \text{SAT} \end{array}$$

■

EXAMPLE 13 Markov Process. Powers of a Matrix. Stochastic Matrix

Suppose that the 2004 state of land use in a city of 60 mi^2 of built-up area is

C: Commercially Used 25% I: Industrially Used 20% R: Residentially Used 55%.

Find the states in 2009, 2014, and 2019, assuming that the transition probabilities for 5-year intervals are given by the matrix **A** and remain practically the same over the time considered.

$$\mathbf{A} = \begin{array}{c} \begin{array}{ccc} \text{From C} & \text{From I} & \text{From R} \end{array} \\ \begin{bmatrix} 0.7 & 0.1 & 0 \\ 0.2 & 0.9 & 0.2 \\ 0.1 & 0 & 0.8 \end{bmatrix} \end{array} \begin{array}{l} \\ \text{To C} \\ \text{To I} \\ \text{To R} \end{array}$$

www.EngineeringEBooksPdf.com

A is a **stochastic matrix**, that is, a square matrix with all entries nonnegative and all column sums equal to 1. Our example concerns a **Markov process**,[1] that is, a process for which the probability of entering a certain state depends only on the last state occupied (and the matrix **A**), not on any earlier state.

Solution. From the matrix **A** and the 2004 state we can compute the 2009 state,

$$\begin{array}{c} \mathrm{C} \\ \mathrm{I} \\ \mathrm{R} \end{array} \begin{bmatrix} 0.7 \cdot 25 + 0.1 \cdot 20 + 0 \cdot 55 \\ 0.2 \cdot 25 + 0.9 \cdot 20 + 0.2 \cdot 55 \\ 0.1 \cdot 25 + 0 \cdot 20 + 0.8 \cdot 55 \end{bmatrix} = \begin{bmatrix} 0.7 & 0.1 & 0 \\ 0.2 & 0.9 & 0.2 \\ 0.1 & 0 & 0.8 \end{bmatrix} \begin{bmatrix} 25 \\ 20 \\ 55 \end{bmatrix} = \begin{bmatrix} 19.5 \\ 34.0 \\ 46.5 \end{bmatrix}.$$

To explain: The 2009 figure for C equals 25% times the probability 0.7 that C goes into C, plus 20% times the probability 0.1 that I goes into C, plus 55% times the probability 0 that R goes into C. Together,

$$25 \cdot 0.7 + 20 \cdot 0.1 + 55 \cdot 0 = 19.5\,[\%]. \qquad \text{Also} \qquad 25 \cdot 0.2 + 20 \cdot 0.9 + 55 \cdot 0.2 = 34\,[\%].$$

Similarly, the new R is 46.5%. We see that the 2009 state vector is the column vector

$$\mathbf{y} = [19.5 \quad 34.0 \quad 46.5]^{\mathsf{T}} = \mathbf{Ax} = \mathbf{A}\,[25 \quad 20 \quad 55]^{\mathsf{T}}$$

where the column vector $\mathbf{x} = [25 \quad 20 \quad 55]^{\mathsf{T}}$ is the given 2004 state vector. Note that the sum of the entries of **y** is 100 [%]. Similarly, you may verify that for 2014 and 2019 we get the state vectors

$$\mathbf{z} = \mathbf{Ay} = \mathbf{A}(\mathbf{Ax}) = \mathbf{A}^2\mathbf{x} = [17.05 \quad 43.80 \quad 39.15]^{\mathsf{T}}$$

$$\mathbf{u} = \mathbf{Az} = \mathbf{A}^2\mathbf{y} = \mathbf{A}^3\mathbf{x} = [16.315 \quad 50.660 \quad 33.025]^{\mathsf{T}}.$$

Answer. In 2009 the commercial area will be 19.5% (11.7 mi^2), the industrial 34% (20.4 mi^2), and the residential 46.5% (27.9 mi^2). For 2014 the corresponding figures are 17.05%, 43.80%, and 39.15%. For 2019 they are 16.315%, 50.660%, and 33.025%. (In Sec. 8.2 we shall see what happens in the limit, assuming that those probabilities remain the same. In the meantime, can you experiment or guess?) ■

PROBLEM SET 7.2

1–10 GENERAL QUESTIONS

1. **Multiplication.** Why is multiplication of matrices restricted by conditions on the factors?
2. **Square matrix.** What form does a 3×3 matrix have if it is symmetric as well as skew-symmetric?
3. **Product of vectors.** Can every 3×3 matrix be represented by two vectors as in Example 3?
4. **Skew-symmetric matrix.** How many different entries can a 4×4 skew-symmetric matrix have? An $n \times n$ skew-symmetric matrix?
5. Same questions as in Prob. 4 for symmetric matrices.
6. **Triangular matrix.** If $\mathbf{U}_1$, $\mathbf{U}_2$ are upper triangular and $\mathbf{L}_1$, $\mathbf{L}_2$ are lower triangular, which of the following are triangular?

$$\mathbf{U}_1 + \mathbf{U}_2, \quad \mathbf{U}_1\mathbf{U}_2, \quad \mathbf{U}_1^2, \quad \mathbf{U}_1 + \mathbf{L}_1, \quad \mathbf{U}_1\mathbf{L}_1, \quad \mathbf{L}_1 + \mathbf{L}_2$$

7. **Idempotent matrix**, defined by $\mathbf{A}^2 = \mathbf{A}$. Can you find four 2×2 idempotent matrices?
8. **Nilpotent matrix**, defined by $\mathbf{B}^m = \mathbf{0}$ for some m. Can you find three 2×2 nilpotent matrices?
9. **Transposition.** Can you prove (10a)–(10c) for 3×3 matrices? For $m \times n$ matrices?
10. **Transposition.** **(a)** Illustrate (10d) by simple examples. **(b)** Prove (10d).

11–20 MULTIPLICATION, ADDITION, AND TRANSPOSITION OF MATRICES AND VECTORS

Let

$$\mathbf{A} = \begin{bmatrix} 4 & -2 & 3 \\ -2 & 1 & 6 \\ 1 & 2 & 2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 & -3 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 \\ 3 & 2 \\ -2 & 0 \end{bmatrix}, \qquad \mathbf{a} = [1 \quad -2 \quad 0], \qquad \mathbf{b} = \begin{bmatrix} 3 \\ 1 \\ -1 \end{bmatrix}.$$

[1] ANDREI ANDREJEVITCH MARKOV (1856–1922), Russian mathematician, known for his work in probability theory.

www.EngineeringEBooksPdf.com

Showing all intermediate results, calculate the following expressions or give reasons why they are undefined:

11. $\mathbf{AB}$, $\mathbf{AB}^\mathsf{T}$, $\mathbf{BA}$, $\mathbf{B}^\mathsf{T}\mathbf{A}$

12. $\mathbf{AA}^\mathsf{T}$, $\mathbf{A}^2$, $\mathbf{BB}^\mathsf{T}$, $\mathbf{B}^2$

13. $\mathbf{CC}^\mathsf{T}$, $\mathbf{BC}$, $\mathbf{CB}$, $\mathbf{C}^\mathsf{T}\mathbf{B}$

14. $3\mathbf{A} - 2\mathbf{B}$, $(3\mathbf{A} - 2\mathbf{B})^\mathsf{T}$, $3\mathbf{A}^\mathsf{T} - 2\mathbf{B}^\mathsf{T}$, $(3\mathbf{A} - 2\mathbf{B})^\mathsf{T}\mathbf{a}^\mathsf{T}$

15. $\mathbf{Aa}$, $\mathbf{Aa}^\mathsf{T}$, $(\mathbf{Ab})^\mathsf{T}$, $\mathbf{b}^\mathsf{T}\mathbf{A}^\mathsf{T}$

16. $\mathbf{BC}$, $\mathbf{BC}^\mathsf{T}$, $\mathbf{Bb}$, $\mathbf{b}^\mathsf{T}\mathbf{B}$

17. $\mathbf{ABC}$, $\mathbf{ABa}$, $\mathbf{ABb}$, $\mathbf{Ca}^\mathsf{T}$

18. $\mathbf{ab}$, $\mathbf{ba}$, $\mathbf{aA}$, $\mathbf{Bb}$

19. $1.5\mathbf{a} + 3.0\mathbf{b}$, $1.5\mathbf{a}^\mathsf{T} + 3.0\mathbf{b}$, $(\mathbf{A} - \mathbf{B})\mathbf{b}$, $\mathbf{Ab} - \mathbf{Bb}$

20. $\mathbf{b}^\mathsf{T}\mathbf{Ab}$, $\mathbf{aBa}^\mathsf{T}$, $\mathbf{aCC}^\mathsf{T}$, $\mathbf{C}^\mathsf{T}\mathbf{ba}$

21. General rules. Prove (2) for 2×2 matrices $\mathbf{A} = [a_{jk}]$, $\mathbf{B} = [b_{jk}]$, $\mathbf{C} = [c_{jk}]$, and a general scalar.

22. Product. Write $\mathbf{AB}$ in Prob. 11 in terms of row and column vectors.

23. Product. Calculate $\mathbf{AB}$ in Prob. 11 columnwise. See Example 1.

24. Commutativity. Find all 2×2 matrices $\mathbf{A} = [a_{jk}]$ that commute with $\mathbf{B} = [b_{jk}]$, where $b_{jk} = j + k$.

25. TEAM PROJECT. Symmetric and Skew-Symmetric Matrices. These matrices occur quite frequently in applications, so it is worthwhile to study some of their most important properties.

(a) Verify the claims in (11) that $a_{kj} = a_{jk}$ for a symmetric matrix, and $a_{kj} = -a_{jk}$ for a skew-symmetric matrix. Give examples.

(b) Show that for every square matrix $\mathbf{C}$ the matrix $\mathbf{C} + \mathbf{C}^\mathsf{T}$ is symmetric and $\mathbf{C} - \mathbf{C}^\mathsf{T}$ is skew-symmetric. Write $\mathbf{C}$ in the form $\mathbf{C} = \mathbf{S} + \mathbf{T}$, where $\mathbf{S}$ is symmetric and $\mathbf{T}$ is skew-symmetric and find $\mathbf{S}$ and $\mathbf{T}$ in terms of $\mathbf{C}$. Represent $\mathbf{A}$ and $\mathbf{B}$ in Probs. 11–20 in this form.

(c) A **linear combination** of matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \cdots, \mathbf{M}$ of the same size is an expression of the form

$$a\mathbf{A} + b\mathbf{B} + c\mathbf{C} + \cdots + m\mathbf{M}, \tag{14}$$

where $a, \cdots, m$ are any scalars. Show that if these matrices are square and symmetric, so is (14); similarly, if they are skew-symmetric, so is (14).

(d) Show that $\mathbf{AB}$ with symmetric $\mathbf{A}$ and $\mathbf{B}$ is symmetric if and only if $\mathbf{A}$ and $\mathbf{B}$ **commute**, that is, $\mathbf{AB} = \mathbf{BA}$.

(e) Under what condition is the product of skew-symmetric matrices skew-symmetric?

26–30 FURTHER APPLICATIONS

26. Production. In a production process, let N mean "no trouble" and T "trouble." Let the transition probabilities from one day to the next be 0.8 for $N \to N$, hence 0.2 for $N \to T$, and 0.5 for $T \to N$, hence 0.5 for $T \to T$. If today there is no trouble, what is the probability of N two days after today? Three days after today?

27. CAS Experiment. Markov Process. Write a program for a Markov process. Use it to calculate further steps in Example 13 of the text. Experiment with other stochastic 3×3 matrices, also using different starting values.

28. Concert subscription. In a community of 100,000 adults, subscribers to a concert series tend to renew their subscription with probability 90% and persons presently not subscribing will subscribe for the next season with probability 0.2%. If the present number of subscribers is 1200, can one predict an increase, decrease, or no change over each of the next three seasons?

29. Profit vector. Two factory outlets F_1 and F_2 in New York and Los Angeles sell sofas (S), chairs (C), and tables (T) with a profit of \$35, \$62, and \$30, respectively. Let the sales in a certain week be given by the matrix

$$\begin{array}{cc} & \begin{array}{ccc} \text{S} & \text{C} & \text{T} \end{array} \\ \mathbf{A} = & \begin{bmatrix} 400 & 60 & 240 \\ 100 & 120 & 500 \end{bmatrix} \begin{array}{c} F_1 \\ F_2 \end{array} \end{array}$$

Introduce a "profit vector" $\mathbf{p}$ such that the components of $\mathbf{v} = \mathbf{Ap}$ give the total profits of F_1 and F_2.

30. TEAM PROJECT. Special Linear Transformations. Rotations have various applications. We show in this project how they can be handled by matrices.

(a) Rotation in the plane. Show that the linear transformation $\mathbf{y} = \mathbf{Ax}$ with

$$\mathbf{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

is a counterclockwise rotation of the Cartesian x_1x_2-coordinate system in the plane about the origin, where θ is the angle of rotation.

(b) Rotation through $n\theta$. Show that in (a)

$$\mathbf{A}^n = \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix}.$$

Is this plausible? Explain this in words.

(c) Addition formulas for cosine and sine. By geometry we should have

$$\begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{bmatrix} = \begin{bmatrix} \cos(\alpha+\beta) & -\sin(\alpha+\beta) \\ \sin(\alpha+\beta) & \cos(\alpha+\beta) \end{bmatrix}.$$

Derive from this the addition formulas (6) in App. A3.1.

www.EngineeringEBooksPdf.com

(d) Computer graphics. To visualize a three-dimensional object with plane faces (e.g., a cube), we may store the position vectors of the vertices with respect to a suitable $x_1x_2x_3$-coordinate system (and a list of the connecting edges) and then obtain a two-dimensional image on a video screen by projecting the object onto a coordinate plane, for instance, onto the x_1x_2-plane by setting $x_3 = 0$. To change the appearance of the image, we can impose a linear transformation on the position vectors stored. Show that a diagonal matrix $\mathbf{D}$ with main diagonal entries 3, 1, $\frac{1}{2}$ gives from an $\mathbf{x} = [x_j]$ the new position vector $\mathbf{y} = \mathbf{Dx}$, where $y_1 = 3x_1$ (stretch in the x_1-direction by a factor 3), $y_2 = x_2$ (unchanged), $y_3 = \frac{1}{2}x_3$ (contraction in the x_3-direction). What effect would a scalar matrix have?

(e) Rotations in space. Explain $\mathbf{y} = \mathbf{Ax}$ geometrically when $\mathbf{A}$ is one of the three matrices

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix},$$

$$\begin{bmatrix} \cos\varphi & 0 & -\sin\varphi \\ 0 & 1 & 0 \\ \sin\varphi & 0 & \cos\varphi \end{bmatrix}, \quad \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

What effect would these transformations have in situations such as that described in (d)?

7.3 Linear Systems of Equations. Gauss Elimination

We now come to one of the most important use of matrices, that is, using matrices to solve systems of linear equations. We showed informally, in Example 1 of Sec. 7.1, how to represent the information contained in a system of linear equations by a matrix, called the augmented matrix. This matrix will then be used in solving the linear system of equations. Our approach to solving linear systems is called the Gauss elimination method. Since this method is so fundamental to linear algebra, the student should be alert.

A shorter term for systems of linear equations is just **linear systems**. Linear systems model many applications in engineering, economics, statistics, and many other areas. Electrical networks, traffic flow, and commodity markets may serve as specific examples of applications.

Linear System, Coefficient Matrix, Augmented Matrix

A **linear system of *m* equations in *n* unknowns** $x_1, \cdots, x_n$ is a set of equations of the form

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m. \end{aligned} \tag{1}$$

The system is called *linear* because each variable x_j appears in the first power only, just as in the equation of a straight line. $a_{11}, \cdots, a_{mn}$ are given numbers, called the **coefficients** of the system. $b_1, \cdots, b_m$ on the right are also given numbers. If all the b_j are zero, then (1) is called a **homogeneous system**. If at least one b_j is not zero, then (1) is called a **nonhomogeneous system**.

www.EngineeringEBooksPdf.com

A **solution** of (1) is a set of numbers $x_1, \cdots, x_n$ that satisfies all the m equations. A **solution vector** of (1) is a vector $\mathbf{x}$ whose components form a solution of (1). If the system (1) is homogeneous, it always has at least the **trivial solution** $x_1 = 0, \cdots, x_n = 0$.

Matrix Form of the Linear System (1). From the definition of matrix multiplication we see that the m equations of (1) may be written as a single vector equation

$$\mathbf{Ax} = \mathbf{b} \tag{2}$$

where the **coefficient matrix** $\mathbf{A} = [a_{jk}]$ is the $m \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

are column vectors. We assume that the coefficients a_{jk} are not all zero, so that $\mathbf{A}$ is not a zero matrix. Note that $\mathbf{x}$ has n components, whereas $\mathbf{b}$ has m components. The matrix

$$\tilde{\mathbf{A}} = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right]$$

is called the **augmented matrix** of the system (1). The dashed vertical line could be omitted, as we shall do later. It is merely a reminder that the last column of $\tilde{\mathbf{A}}$ did not come from matrix $\mathbf{A}$ but came from vector $\mathbf{b}$. Thus, we *augmented* the matrix $\mathbf{A}$.

Note that ***the augmented matrix*** $\tilde{\mathbf{A}}$ ***determines the system*** **(1)** ***completely*** because it contains all the given numbers appearing in (1).

EXAMPLE 1 Geometric Interpretation. Existence and Uniqueness of Solutions

If $m = n = 2$, we have two equations in two unknowns x_1, x_2

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2.$$

If we interpret x_1, x_2 as coordinates in the x_1x_2-plane, then each of the two equations represents a straight line, and (x_1, x_2) is a solution if and only if the point P with coordinates x_1, x_2 lies on both lines. Hence there are three possible cases (see Fig. 158 on next page):

(*a*) Precisely one solution if the lines intersect

(*b*) Infinitely many solutions if the lines coincide

(*c*) No solution if the lines are parallel

www.EngineeringEBooksPdf.com

Fig. 158. Three equations in three unknowns interpreted as planes in space

For instance,

If the system is homogenous, Case (c) cannot happen, because then those two straight lines pass through the origin, whose coordinates (0, 0) constitute the trivial solution. Similarly, our present discussion can be extended from two equations in two unknowns to three equations in three unknowns. We give the geometric interpretation of three possible cases concerning solutions in Fig. 158. Instead of straight lines we have planes and the solution depends on the positioning of these planes in space relative to each other. The student may wish to come up with some specific examples. ■

Our simple example illustrated that a system (1) may have no solution. This leads to such questions as: Does a given system (1) have a solution? Under what conditions does it have precisely one solution? If it has more than one solution, how can we characterize the set of all solutions? We shall consider such questions in Sec. 7.5.

First, however, let us discuss an important systematic method for solving linear systems.

Gauss Elimination and Back Substitution

The Gauss elimination method can be motivated as follows. Consider a linear system that is in ***triangular form*** (in full, *upper* triangular form) such as

$$\begin{aligned} 2x_1 + 5x_2 &= 2 \\ 13x_2 &= -26 \end{aligned}$$

(*Triangular* means that all the nonzero entries of the corresponding coefficient matrix lie above the diagonal and form an upside-down 90° triangle.) Then we can solve the system by **back substitution**, that is, we solve the last equation for the variable, $x_2 = -26/13 = -2$, and then work backward, substituting $x_2 = -2$ into the first equation and solving it for x_1, obtaining $x_1 = \frac{1}{2}(2 - 5x_2) = \frac{1}{2}(2 - 5 \cdot (-2)) = 6$. This gives us the idea of first reducing a general system to triangular form. For instance, let the given system be

$$\begin{aligned} 2x_1 + 5x_2 &= 2 \\ -4x_1 + 3x_2 &= -30. \end{aligned} \qquad \text{Its augmented matrix is} \qquad \begin{bmatrix} 2 & 5 & 2 \\ -4 & 3 & -30 \end{bmatrix}.$$

We leave the first equation as it is. We eliminate x_1 from the second equation, to get a triangular system. For this we add twice the first equation to the second, and we do the same

www.EngineeringEBooksPdf.com

operation on the ***rows*** of the augmented matrix. This gives $-4x_1 + 4x_1 + 3x_2 + 10x_2 = -30 + 2 \cdot 2$, that is,

$$\begin{aligned} 2x_1 + 5x_2 &= 2 \\ 13x_2 &= -26 \end{aligned} \qquad \text{Row 2} + 2\,\text{Row 1} \qquad \begin{bmatrix} 2 & 5 & 2 \\ 0 & 13 & -26 \end{bmatrix}$$

where Row 2 + 2 Row 1 means "Add twice Row 1 to Row 2" in the original matrix. This is the **Gauss elimination** (for 2 equations in 2 unknowns) giving the triangular form, from which back substitution now yields $x_2 = -2$ and $x_1 = 6$, as before.

Since a linear system is completely determined by its augmented matrix, ***Gauss elimination can be done by merely considering the matrices***, as we have just indicated. We do this again in the next example, emphasizing the matrices by writing them first and the equations behind them, just as a help in order not to lose track.

EXAMPLE 2 Gauss Elimination. Electrical Network

Solve the linear system

$$\begin{aligned} x_1 - x_2 + x_3 &= 0 \\ -x_1 + x_2 - x_3 &= 0 \\ 10x_2 + 25x_3 &= 90 \\ 20x_1 + 10x_2 &= 80. \end{aligned}$$

Derivation from the circuit in Fig. 159 (Optional). This is the system for the unknown currents $x_1 = i_1$, $x_2 = i_2$, $x_3 = i_3$ in the electrical network in Fig. 159. To obtain it, we label the currents as shown, choosing directions arbitrarily; if a current will come out negative, this will simply mean that the current flows against the direction of our arrow. The current entering each battery will be the same as the current leaving it. The equations for the currents result from Kirchhoff's laws:

Kirchhoff's Current Law (KCL). *At any point of a circuit, the sum of the inflowing currents equals the sum of the outflowing currents.*

Kirchhoff's Voltage Law (KVL). *In any closed loop, the sum of all voltage drops equals the impressed electromotive force.*

Node P gives the first equation, node Q the second, the right loop the third, and the left loop the fourth, as indicated in the figure.

Node P: $i_1 - i_2 + i_3 = 0$

Node Q: $-i_1 + i_2 - i_3 = 0$

Right loop: $10i_2 + 25i_3 = 90$

Left loop: $20i_1 + 10i_2 = 80$

Fig. 159. Network in Example 2 and equations relating the currents

Solution by Gauss Elimination. This system could be solved rather quickly by noticing its particular form. But this is not the point. The point is that the Gauss elimination is systematic and will work in general,

www.EngineeringEBooksPdf.com

also for large systems. We apply it to our system and then do back substitution. As indicated, let us write the augmented matrix of the system first and then the system itself:

$$\begin{array}{l} \\ \text{Pivot 1} \longrightarrow \\ \\ \text{Eliminate} \longrightarrow \\ \\ \end{array} \begin{array}{c} \text{Augmented Matrix } \tilde{\mathbf{A}} \\ \left[\begin{array}{rrr|r} 1 & -1 & 1 & 0 \\ -1 & 1 & -1 & 0 \\ 0 & 10 & 25 & 90 \\ 20 & 10 & 0 & 80 \end{array}\right] \end{array} \qquad \begin{array}{l} \\ \text{Pivot 1} \longrightarrow \\ \\ \text{Eliminate} \longrightarrow \\ \\ \end{array} \begin{array}{c} \text{Equations} \\ \begin{array}{rrrl} x_1 - & x_2 + & x_3 = & 0 \\ -x_1 + & x_2 - & x_3 = & 0 \\ & 10x_2 + & 25x_3 = & 90 \\ 20x_1 + & 10x_2 & = & 80. \end{array} \end{array}$$

Step 1. Elimination of x_1

Call the first row of **A** the **pivot row** and the first equation the **pivot equation**. Call the coefficient 1 of its x_1-term the **pivot** in this step. Use this equation to eliminate x_1 (get rid of x_1) in the other equations. For this, do:

Add 1 times the pivot equation to the second equation.

Add -20 times the pivot equation to the fourth equation.

This corresponds to **row operations** on the augmented matrix as indicated in BLUE behind the ***new matrix*** in (3). So the operations are performed on the ***preceding matrix***. The result is

$$(3) \qquad \left[\begin{array}{rrr|r} 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 10 & 25 & 90 \\ 0 & 30 & -20 & 80 \end{array}\right] \begin{array}{l} \\ \text{Row 2 + Row 1} \\ \\ \text{Row 4 − 20 Row 1} \end{array} \qquad \begin{array}{rrrl} x_1 - & x_2 + & x_3 = & 0 \\ & & 0 = & 0 \\ & 10x_2 + & 25x_3 = & 90 \\ & 30x_2 - & 20x_3 = & 80. \end{array}$$

Step 2. Elimination of x_2

The first equation remains as it is. We want the new second equation to serve as the next pivot equation. But since it has no x_2-term (in fact, it is $0 = 0$), we must first change the order of the equations and the corresponding rows of the new matrix. We put $0 = 0$ at the end and move the third equation and the fourth equation one place up. This is called **partial pivoting** (as opposed to the rarely used *total pivoting*, in which the order of the unknowns is also changed). It gives

$$\begin{array}{l} \\ \text{Pivot 10} \longrightarrow \\ \text{Eliminate 30} \longrightarrow \\ \\ \end{array} \left[\begin{array}{rrr|r} 1 & -1 & 1 & 0 \\ 0 & 10 & 25 & 90 \\ 0 & 30 & -20 & 80 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{l} \\ \text{Pivot 10} \longrightarrow \\ \text{Eliminate } 30x_2 \longrightarrow \\ \\ \end{array} \begin{array}{rrrl} x_1 - & x_2 + & x_3 = & 0 \\ & 10x_2 + & 25x_3 = & 90 \\ & 30x_2 - & 20x_3 = & 80 \\ & & 0 = & 0. \end{array}$$

To eliminate x_2, do:

Add -3 times the pivot equation to the third equation.

The result is

$$(4) \qquad \left[\begin{array}{rrr|r} 1 & -1 & 1 & 0 \\ 0 & 10 & 25 & 90 \\ 0 & 0 & -95 & -190 \\ 0 & 0 & 0 & 0 \end{array}\right] \begin{array}{l} \\ \\ \text{Row 3 − 3 Row 2} \\ \\ \end{array} \qquad \begin{array}{rrrl} x_1 - & x_2 + & x_3 = & 0 \\ & 10x_2 + & 25x_3 = & 90 \\ & & -95x_3 = & -190 \\ & & 0 = & 0. \end{array}$$

Back Substitution. ***Determination of x_3, x_2, x_1 (in this order)***

Working backward from the last to the first equation of this "triangular" system (4), we can now readily find x_3, then x_2, and then x_1:

$$\begin{aligned} -95x_3 &= -190 & x_3 &= i_3 = 2\,[\mathrm{A}] \\ 10x_2 + 25x_3 &= 90 & x_2 &= \tfrac{1}{10}(90 - 25x_3) = i_2 = 4\,[\mathrm{A}] \\ x_1 - x_2 + x_3 &= 0 & x_1 &= x_2 - x_3 = i_1 = 2\,[\mathrm{A}] \end{aligned}$$

where A stands for "amperes." This is the answer to our problem. The solution is unique. ■

www.EngineeringEBooksPdf.com

Elementary Row Operations. Row-Equivalent Systems

Example 2 illustrates the operations of the Gauss elimination. These are the first two of three operations, which are called

Elementary Row Operations for Matrices:

Interchange of two rows

Addition of a constant multiple of one row to another row

Multiplication of a row by a ***nonzero*** *constant c*

CAUTION! These operations are for rows, ***not for columns***! They correspond to the following

Elementary Operations for Equations:

Interchange of two equations

Addition of a constant multiple of one equation to another equation

Multiplication of an equation by a ***nonzero*** *constant c*

Clearly, the interchange of two equations does not alter the solution set. Neither does their addition because we can undo it by a corresponding subtraction. Similarly for their multiplication, which we can undo by multiplying the new equation by $1/c$ (since $c \neq 0$), producing the original equation.

We now call a linear system S_1 **row-equivalent** to a linear system S_2 if S_1 can be obtained from S_2 by (finitely many!) row operations. This justifies Gauss elimination and establishes the following result.

THEOREM 1

Row-Equivalent Systems

Row-equivalent linear systems have the same set of solutions.

Because of this theorem, systems having the same solution sets are often called *equivalent systems.* But note well that we are dealing with ***row operations***. No column operations on the augmented matrix are permitted in this context because they would generally alter the solution set.

A linear system (1) is called **overdetermined** if it has more equations than unknowns, as in Example 2, **determined** if $m = n$, as in Example 1, and **underdetermined** if it has fewer equations than unknowns.

Furthermore, a system (1) is called **consistent** if it has at least one solution (thus, one solution or infinitely many solutions), but **inconsistent** if it has no solutions at all, as $x_1 + x_2 = 1, x_1 + x_2 = 0$ in Example 1, Case (c).

Gauss Elimination: The Three Possible Cases of Systems

We have seen, in Example 2, that Gauss elimination can solve linear systems that have a unique solution. This leaves us to apply Gauss elimination to a system with infinitely many solutions (in Example 3) and one with no solution (in Example 4).

www.EngineeringEBooksPdf.com

EXAMPLE 3 **Gauss Elimination if Infinitely Many Solutions Exist**

Solve the following linear system of three equations in four unknowns whose augmented matrix is

$$(5)\qquad \left[\begin{array}{cccc|c} 3.0 & 2.0 & 2.0 & -5.0 & 8.0 \\ 0.6 & 1.5 & 1.5 & -5.4 & 2.7 \\ 1.2 & -0.3 & -0.3 & 2.4 & 2.1 \end{array}\right]. \qquad \text{Thus,} \qquad \begin{aligned} 3.0x_1 + 2.0x_2 + 2.0x_3 - 5.0x_4 &= 8.0 \\ \boxed{0.6x_1} + 1.5x_2 + 1.5x_3 - 5.4x_4 &= 2.7 \\ \boxed{1.2x_1} - 0.3x_2 - 0.3x_3 + 2.4x_4 &= 2.1. \end{aligned}$$

Solution. As in the previous example, we circle pivots and box terms of equations and corresponding entries to be eliminated. We indicate the operations in terms of equations and operate on both equations and matrices.

Step 1. Elimination of x_1 from the second and third equations by adding

$$-0.6/3.0 = -0.2 \text{ times the first equation to the second equation,}$$

$$-1.2/3.0 = -0.4 \text{ times the first equation to the third equation.}$$

This gives the following, in which the pivot of the next step is circled.

$$(6)\qquad \left[\begin{array}{cccc|c} 3.0 & 2.0 & 2.0 & -5.0 & 8.0 \\ 0 & 1.1 & 1.1 & -4.4 & 1.1 \\ 0 & -1.1 & -1.1 & 4.4 & -1.1 \end{array}\right] \begin{array}{l} \\ \text{Row 2} - 0.2 \text{ Row 1} \\ \text{Row 3} - 0.4 \text{ Row 1} \end{array} \qquad \begin{aligned} 3.0x_1 + 2.0x_2 + 2.0x_3 - 5.0x_4 &= 8.0 \\ 1.1x_2 + 1.1x_3 - 4.4x_4 &= 1.1 \\ \boxed{-1.1x_2} - 1.1x_3 + 4.4x_4 &= -1.1. \end{aligned}$$

Step 2. Elimination of x_2 from the third equation of (6) by adding

$$1.1/1.1 = 1 \text{ times the second equation to the third equation.}$$

This gives

$$(7)\qquad \left[\begin{array}{cccc|c} 3.0 & 2.0 & 2.0 & -5.0 & 8.0 \\ 0 & 1.1 & 1.1 & -4.4 & 1.1 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \begin{array}{l} \\ \\ \text{Row 3} + \text{Row 2} \end{array} \qquad \begin{aligned} 3.0x_1 + 2.0x_2 + 2.0x_3 - 5.0x_4 &= 8.0 \\ 1.1x_2 + 1.1x_3 - 4.4x_4 &= 1.1 \\ 0 &= 0. \end{aligned}$$

Back Substitution. From the second equation, $x_2 = 1 - x_3 + 4x_4$. From this and the first equation, $x_1 = 2 - x_4$. Since x_3 and x_4 remain arbitrary, we have infinitely many solutions. If we choose a value of x_3 and a value of x_4, then the corresponding values of x_1 and x_2 are uniquely determined.

On Notation. If unknowns remain arbitrary, it is also customary to denote them by other letters $t_1, t_2, \cdots$. In this example we may thus write $x_1 = 2 - x_4 = 2 - t_2$, $x_2 = 1 - x_3 + 4x_4 = 1 - t_1 + 4t_2$, $x_3 = t_1$ (first arbitrary unknown), $x_4 = t_2$ (second arbitrary unknown). ■

EXAMPLE 4 **Gauss Elimination if no Solution Exists**

What will happen if we apply the Gauss elimination to a linear system that has no solution? The answer is that in this case the method will show this fact by producing a contradiction. For instance, consider

$$\left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 2 & 1 & 1 & 0 \\ 6 & 2 & 4 & 6 \end{array}\right] \qquad \begin{aligned} 3x_1 + 2x_2 + x_3 &= 3 \\ \boxed{2x_1} + x_2 + x_3 &= 0 \\ \boxed{6x_1} + 2x_2 + 4x_3 &= 6. \end{aligned}$$

Step 1. Elimination of x_1 from the second and third equations by adding

$$-\tfrac{2}{3} \text{ times the first equation to the second equation,}$$

$$-\tfrac{6}{3} = -2 \text{ times the first equation to the third equation.}$$

www.EngineeringEBooksPdf.com

This gives

$$\left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -\frac{1}{3} & \frac{1}{3} & -2 \\ 0 & -2 & 2 & 0 \end{array}\right] \begin{array}{l} \\ \text{Row 2} - \frac{2}{3}\text{ Row 1} \\ \text{Row 3} - 2\text{ Row 1} \end{array} \qquad \begin{array}{rcl} 3x_1 + 2x_2 + x_3 &=& 3 \\ -\frac{1}{3}x_2 + \frac{1}{3}x_3 &=& -2 \\ \boxed{-2x_2} + 2x_3 &=& 0. \end{array}$$

Step 2. Elimination of x_2 from the third equation gives

$$\left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -\frac{1}{3} & \frac{1}{3} & -2 \\ 0 & 0 & 0 & 12 \end{array}\right] \begin{array}{l} \\ \\ \text{Row 3} - 6\text{ Row 2} \end{array} \qquad \begin{array}{rcl} 3x_1 + 2x_2 + x_3 &=& 3 \\ -\frac{1}{3}x_2 + \frac{1}{3}x_3 &=& -2 \\ 0 &=& 12. \end{array}$$

The false statement $0 = 12$ shows that the system has no solution. ■

Row Echelon Form and Information From It

At the end of the Gauss elimination the form of the coefficient matrix, the augmented matrix, and the system itself are called the **row echelon form**. In it, rows of zeros, if present, are the last rows, and, in each nonzero row, the leftmost nonzero entry is farther to the right than in the previous row. For instance, in Example 4 the coefficient matrix and its augmented in row echelon form are

$$\begin{bmatrix} 3 & 2 & 1 \\ 0 & -\frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -\frac{1}{3} & \frac{1}{3} & -2 \\ 0 & 0 & 0 & 12 \end{array}\right]. \tag{8}$$

Note that we do not require that the leftmost nonzero entries be 1 since this would have no theoretic or numeric advantage. (The so-called *reduced echelon form*, in which those entries *are* 1, will be discussed in Sec. 7.8.)

The original system of m equations in n unknowns has augmented matrix $[\mathbf{A}|\mathbf{b}]$. This is to be row reduced to matrix $[\mathbf{R}|\mathbf{f}]$. The two systems $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{Rx} = \mathbf{f}$ are equivalent: if either one has a solution, so does the other, and the solutions are identical.

At the end of the Gauss elimination (before the back substitution), the row echelon form of the augmented matrix will be

$$\left[\begin{array}{cccc|c} r_{11} & r_{12} & \cdots\cdots & r_{1n} & f_1 \\ & r_{22} & \cdots\cdots & r_{2n} & f_2 \\ & & \ddots & \vdots & \vdots \\ & & r_{rr} \;\cdots & r_{rn} & f_r \\ & & & & f_{r+1} \\ & & & & \vdots \\ & & & & f_m \end{array}\right]. \tag{9}$$

Here, $r \leqq m$, $r_{11} \neq 0$, and all entries in the blue triangle and blue rectangle are zero.

The number of nonzero rows, r, in the row-reduced coefficient matrix $\mathbf{R}$ is called the **rank of R** and also the **rank of A**. Here is the method for determining whether $\mathbf{Ax} = \mathbf{b}$ has solutions and what they are:

(a) No solution. If r is less than m (meaning that $\mathbf{R}$ actually has at least one row of all 0s) *and* at least one of the numbers $f_{r+1}, f_{r+2}, \cdots, f_m$ is not zero, then the system

www.EngineeringEBooksPdf.com

$\mathbf{Rx} = \mathbf{f}$ is inconsistent: No solution is possible. Therefore the system $\mathbf{Ax} = \mathbf{b}$ is inconsistent as well. See Example 4, where $r = 2 < m = 3$ and $f_{r+1} = f_3 = 12$.

If the system is consistent (either $r = m$, or $r < m$ *and* all the numbers $f_{r+1}, f_{r+2}, \cdots, f_m$ are zero), then there are solutions.

(b) Unique solution. If the system is consistent and $r = n$, there is exactly one solution, which can be found by back substitution. See Example 2, where $r = n = 3$ and $m = 4$.

(c) Infinitely many solutions. To obtain any of these solutions, choose values of $x_{r+1}, \cdots, x_n$ arbitrarily. Then solve the rth equation for x_r (in terms of those arbitrary values), then the $(r-1)$st equation for x_{r-1}, and so on up the line. See Example 3.

Orientation. Gauss elimination is reasonable in computing time and storage demand. We shall consider those aspects in Sec. 20.1 in the chapter on numeric linear algebra. Section 7.4 develops fundamental concepts of linear algebra such as linear independence and rank of a matrix. These in turn will be used in Sec. 7.5 to fully characterize the behavior of linear systems in terms of existence and uniqueness of solutions.

PROBLEM SET 7.3

1–14 GAUSS ELIMINATION

Solve the linear system given explicitly or by its augmented matrix. Show details.

1. $\begin{aligned} 4x - 6y &= -11 \\ -3x + 8y &= 10 \end{aligned}$

2. $\begin{bmatrix} 3.0 & -0.5 & 0.6 \\ 1.5 & 4.5 & 6.0 \end{bmatrix}$

3. $\begin{aligned} x + y - z &= 9 \\ 8y + 6z &= -6 \\ -2x + 4y - 6z &= 40 \end{aligned}$

4. $\begin{bmatrix} 4 & 1 & 0 & 4 \\ 5 & -3 & 1 & 2 \\ -9 & 2 & -1 & 5 \end{bmatrix}$

5. $\begin{bmatrix} 13 & 12 & -6 \\ -4 & 7 & -73 \\ 11 & -13 & 157 \end{bmatrix}$

6. $\begin{bmatrix} 4 & -8 & 3 & 16 \\ -1 & 2 & -5 & -21 \\ 3 & -6 & 1 & 7 \end{bmatrix}$

7. $\begin{bmatrix} 2 & 4 & 1 & 0 \\ -1 & 1 & -2 & 0 \\ 4 & 0 & 6 & 0 \end{bmatrix}$

8. $\begin{aligned} 4y + 3z &= 8 \\ 2x \qquad - z &= 2 \\ 3x + 2y \qquad &= 5 \end{aligned}$

9. $\begin{aligned} -2y - 2z &= -8 \\ 3x + 4y - 5z &= 13 \end{aligned}$

10. $\begin{bmatrix} 5 & -7 & 3 & 17 \\ -15 & 21 & -9 & 50 \end{bmatrix}$

11. $\begin{bmatrix} 0 & 5 & 5 & -10 & 0 \\ 2 & -3 & -3 & 6 & 2 \\ 4 & 1 & 1 & -2 & 4 \end{bmatrix}$

12. $\begin{bmatrix} 2 & -2 & 4 & 0 & 0 \\ -3 & 3 & -6 & 5 & 15 \\ 1 & -1 & 2 & 0 & 0 \end{bmatrix}$

13. $\begin{aligned} 10x + 4y - 2z &= -4 \\ -3w - 17x + y + 2z &= 2 \\ w + x + y \qquad &= 6 \\ 8w - 34x + 16y - 10z &= 4 \end{aligned}$

14. $\begin{bmatrix} 2 & 3 & 1 & -11 & 1 \\ 5 & -2 & 5 & -4 & 5 \\ 1 & -1 & 3 & -3 & 3 \\ 3 & 4 & -7 & 2 & -7 \end{bmatrix}$

15. Equivalence relation. By definition, an *equivalence relation* on a set is a relation satisfying three conditions: (named as indicated)

(i) Each element A of the set is equivalent to itself (*Reflexivity*).

(ii) If A is equivalent to B, then B is equivalent to A (*Symmetry*).

(iii) If A is equivalent to B and B is equivalent to C, then A is equivalent to C (*Transitivity*).

Show that row equivalence of matrices satisfies these three conditions. *Hint*. Show that for each of the three elementary row operations these conditions hold.

www.EngineeringEBooksPdf.com

16. CAS PROJECT. Gauss Elimination and Back Substitution. Write a program for Gauss elimination and back substitution **(a)** that does not include pivoting and **(b)** that does include pivoting. Apply the programs to Probs. 11–14 and to some larger systems of your choice.

17–21 MODELS OF NETWORKS

In Probs. 17–19, using Kirchhoff's laws (see Example 2) and showing the details, find the currents:

17.

18.

19.

20. Wheatstone bridge. Show that if $R_x/R_3 = R_1/R_2$ in the figure, then $I = 0$. (R_0 is the resistance of the instrument by which I is measured.) This bridge is a method for determining R_x. R_1, R_2, R_3 are known. R_3 is variable. To get R_x, make $I = 0$ by varying R_3. Then calculate $R_x = R_3R_1/R_2$.

Wheatstone bridge

Problem 20

Net of one-way streets

Problem 21

21. Traffic flow. Methods of electrical circuit analysis have applications to other fields. For instance, applying the analog of Kirchhoff's Current Law, find the traffic flow (cars per hour) in the net of one-way streets (in the directions indicated by the arrows) shown in the figure. Is the solution unique?

22. Models of markets. Determine the equilibrium solution ($D_1 = S_1, D_2 = S_2$) of the two-commodity market with linear model (D, S, P = demand, supply, price; index 1 = first commodity, index 2 = second commodity)

$$D_1 = 40 - 2P_1 - P_2, \qquad S_1 = 4P_1 - P_2 + 4,$$

$$D_2 = 5P_1 - 2P_2 + 16, \qquad S_2 = 3P_2 - 4.$$

23. Balancing a chemical equation $x_1C_3H_8 + x_2O_2 \rightarrow x_3CO_2 + x_4H_2O$ means finding integer x_1, x_2, x_3, x_4 such that the numbers of atoms of carbon (C), hydrogen (H), and oxygen (O) are the same on both sides of this reaction, in which propane C_3H_8 and O_2 give carbon dioxide and water. Find the smallest positive integers $x_1, \cdots, x_4$.

24. PROJECT. Elementary Matrices. The idea is that elementary operations can be accomplished by matrix multiplication. If $\mathbf{A}$ is an $m \times n$ matrix on which we want to do an elementary operation, then there is a matrix $\mathbf{E}$ such that $\mathbf{EA}$ is the new matrix after the operation. Such an $\mathbf{E}$ is called an **elementary matrix**. This idea can be helpful, for instance, in the design of algorithms. (*Computationally*, it is generally preferable to do row operations *directly*, rather than by multiplication by $\mathbf{E}$.)

(a) Show that the following are elementary matrices, for interchanging Rows 2 and 3, for adding -5 times the first row to the third, and for multiplying the fourth row by 8.

$$\mathbf{E}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{E}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{E}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}.$$

www.EngineeringEBooksPdf.com

Apply $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3$ to a vector and to a 4×3 matrix of your choice. Find $\mathbf{B} = \mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{A}$, where $\mathbf{A} = [a_{jk}]$ is the general 4×2 matrix. Is $\mathbf{B}$ equal to $\mathbf{C} = \mathbf{E}_1\mathbf{E}_2\mathbf{E}_3\mathbf{A}$?

(b) Conclude that $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3$ are obtained by doing the corresponding elementary operations on the 4×4 unit matrix. Prove that *if* $\mathbf{M}$ *is obtained from* $\mathbf{A}$ *by an elementary row operation, then*

$$\mathbf{M} = \mathbf{EA},$$

where $\mathbf{E}$ *is obtained from the* $n \times n$ *unit matrix* $\mathbf{I}_n$ *by the same row operation.*

7.4 Linear Independence. Rank of a Matrix. Vector Space

Since our next goal is to fully characterize the behavior of linear systems in terms of existence and uniqueness of solutions (Sec. 7.5), we have to introduce new fundamental linear algebraic concepts that will aid us in doing so. Foremost among these are ***linear independence*** and the ***rank of a matrix***. Keep in mind that these concepts are intimately linked with the important Gauss elimination method and how it works.

Linear Independence and Dependence of Vectors

Given any set of m vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(m)}$ (with the same number of components), a **linear combination** of these vectors is an expression of the form

$$c_1\mathbf{a}_{(1)} + c_2\mathbf{a}_{(2)} + \cdots + c_m\mathbf{a}_{(m)}$$

where $c_1, c_2, \cdots, c_m$ are any scalars. Now consider the equation

$$c_1\mathbf{a}_{(1)} + c_2\mathbf{a}_{(2)} + \cdots + c_m\mathbf{a}_{(m)} = \mathbf{0}. \tag{1}$$

Clearly, this vector equation (1) holds if we choose all c_j's zero, because then it becomes $\mathbf{0} = \mathbf{0}$. If this is the only m-tuple of scalars for which (1) holds, then our vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(m)}$ are said to form a *linearly independent set* or, more briefly, we call them **linearly independent**. Otherwise, if (1) also holds with scalars not all zero, we call these vectors **linearly dependent**. This means that we can express at least one of the vectors as a linear combination of the other vectors. For instance, if (1) holds with, say, $c_1 \neq 0$, we can solve (1) for $\mathbf{a}_{(1)}$:

$$\mathbf{a}_{(1)} = k_2\mathbf{a}_{(2)} + \cdots + k_m\mathbf{a}_{(m)} \qquad \text{where } k_j = -c_j/c_1.$$

(Some k_j's may be zero. Or even all of them, namely, if $\mathbf{a}_{(1)} = \mathbf{0}$.)

Why is linear independence important? Well, if a set of vectors is linearly dependent, then we can get rid of at least one or perhaps more of the vectors until we get a linearly independent set. This set is then the smallest "truly essential" set with which we can work. Thus, we cannot express any of the vectors, of this set, linearly in terms of the others.

www.EngineeringEBooksPdf.com

EXAMPLE 1 Linear Independence and Dependence

The three vectors

$$\mathbf{a}_{(1)} = [\ 3 \quad 0 \quad 2 \quad 2]$$
$$\mathbf{a}_{(2)} = [-6 \quad 42 \quad 24 \quad 54]$$
$$\mathbf{a}_{(3)} = [\ 21 \quad -21 \quad 0 \quad -15]$$

are linearly dependent because

$$6\mathbf{a}_{(1)} - \tfrac{1}{2}\mathbf{a}_{(2)} - \mathbf{a}_{(3)} = \mathbf{0}.$$

Although this is easily checked by vector arithmetic (do it!), it is not so easy to discover. However, a systematic method for finding out about linear independence and dependence follows below.

The first two of the three vectors are linearly independent because $c_1\mathbf{a}_{(1)} + c_2\mathbf{a}_{(2)} = \mathbf{0}$ implies $c_2 = 0$ (from the second components) and then $c_1 = 0$ (from any other component of $\mathbf{a}_{(1)}$. ■

Rank of a Matrix

DEFINITION

The **rank** of a matrix $\mathbf{A}$ is the maximum number of linearly independent row vectors of $\mathbf{A}$. It is denoted by rank $\mathbf{A}$.

Our further discussion will show that the rank of a matrix is an important key concept for understanding general properties of matrices and linear systems of equations.

EXAMPLE 2 Rank

The matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 2 & 2 \\ -6 & 42 & 24 & 54 \\ 21 & -21 & 0 & -15 \end{bmatrix} \tag{2}$$

has rank 2, because Example 1 shows that the first two row vectors are linearly independent, whereas all three row vectors are linearly dependent.

Note further that rank $\mathbf{A} = 0$ if and only if $\mathbf{A} = 0$. This follows directly from the definition. ■

We call a matrix $\mathbf{A}_1$ **row-equivalent** to a matrix $\mathbf{A}_2$ if $\mathbf{A}_1$ can be obtained from $\mathbf{A}_2$ by (finitely many!) elementary row operations.

Now the maximum number of linearly independent row vectors of a matrix does not change if we change the order of rows or multiply a row by a nonzero c or take a linear combination by adding a multiple of a row to another row. This shows that rank is **invariant** under elementary row operations:

THEOREM 1 Row-Equivalent Matrices

Row-equivalent matrices have the same rank.

Hence we can determine the rank of a matrix by reducing the matrix to row-echelon form, as was done in Sec. 7.3. Once the matrix is in row-echelon form, we count the number of nonzero rows, which is precisely the rank of the matrix.

www.EngineeringEBooksPdf.com

EXAMPLE 3 Determination of Rank

For the matrix in Example 2 we obtain successively

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 2 & 2 \\ -6 & 42 & 24 & 54 \\ 21 & -21 & 0 & -15 \end{bmatrix} \quad \text{(given)}$$

$$\begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 42 & 28 & 58 \\ 0 & -21 & -14 & -29 \end{bmatrix} \begin{matrix} \\ \text{Row 2} + 2\,\text{Row 1} \\ \text{Row 3} - 7\,\text{Row 1} \end{matrix}$$

$$\begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 42 & 28 & 58 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \\ \\ \text{Row 3} + \frac{1}{2}\,\text{Row 2.} \end{matrix}$$

The last matrix is in row-echelon form and has two nonzero rows. Hence rank $\mathbf{A} = 2$, as before. ■

Examples 1–3 illustrate the following useful theorem (with $p = 3$, $n = 3$, and the rank of the matrix $= 2$).

THEOREM 2 Linear Independence and Dependence of Vectors

Consider p vectors that each have n components. Then these vectors are linearly independent if the matrix formed, with these vectors as row vectors, has rank p. However, these vectors are linearly dependent if that matrix has rank less than p.

Further important properties will result from the basic

THEOREM 3 Rank in Terms of Column Vectors

The rank r of a matrix $\mathbf{A}$ equals the maximum number of linearly independent ***column*** *vectors of $\mathbf{A}$.*

Hence $\mathbf{A}$ and its transpose $\mathbf{A}^{\mathsf{T}}$ have the same rank.

PROOF In this proof we write simply "rows" and "columns" for row and column vectors. Let $\mathbf{A}$ be an $m \times n$ matrix of rank $\mathbf{A} = r$. Then by definition of rank, $\mathbf{A}$ has r linearly independent rows which we denote by $\mathbf{v}_{(1)}, \cdots, \mathbf{v}_{(r)}$ (regardless of their position in $\mathbf{A}$), and all the rows $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(m)}$ of $\mathbf{A}$ are linear combinations of those, say,

$$\begin{aligned} \mathbf{a}_{(1)} &= c_{11}\mathbf{v}_{(1)} + c_{12}\mathbf{v}_{(2)} + \cdots + c_{1r}\mathbf{v}_{(r)} \\ \mathbf{a}_{(2)} &= c_{21}\mathbf{v}_{(1)} + c_{22}\mathbf{v}_{(2)} + \cdots + c_{2r}\mathbf{v}_{(r)} \\ &\vdots \\ \mathbf{a}_{(m)} &= c_{m1}\mathbf{v}_{(1)} + c_{m2}\mathbf{v}_{(2)} + \cdots + c_{mr}\mathbf{v}_{(r)}. \end{aligned} \tag{3}$$

www.EngineeringEBooksPdf.com

These are vector equations for rows. To switch to columns, we write (3) in terms of components as n such systems, with $k = 1, \cdots, n$,

$$
\begin{aligned}
a_{1k} &= c_{11}v_{1k} + c_{12}v_{2k} + \cdots + c_{1r}v_{rk} \\
a_{2k} &= c_{21}v_{1k} + c_{22}v_{2k} + \cdots + c_{2r}v_{rk} \\
&\vdots \\
a_{mk} &= c_{m1}v_{1k} + c_{m2}v_{2k} + \cdots + c_{mr}v_{rk}
\end{aligned}
\tag{4}
$$

and collect components in columns. Indeed, we can write (4) as

$$
\begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{mk} \end{bmatrix} = v_{1k} \begin{bmatrix} c_{11} \\ c_{21} \\ \vdots \\ c_{m1} \end{bmatrix} + v_{2k} \begin{bmatrix} c_{12} \\ c_{22} \\ \vdots \\ c_{m2} \end{bmatrix} + \cdots + v_{rk} \begin{bmatrix} c_{1r} \\ c_{2r} \\ \vdots \\ c_{mr} \end{bmatrix}
\tag{5}
$$

where $k = 1, \cdots, n$. Now the vector on the left is the kth column vector of $\mathbf{A}$. We see that each of these n columns is a linear combination of the same r columns on the right. Hence $\mathbf{A}$ cannot have more linearly independent columns than rows, whose number is rank $\mathbf{A} = r$. Now rows of $\mathbf{A}$ are columns of the transpose $\mathbf{A}^{\mathsf{T}}$. For $\mathbf{A}^{\mathsf{T}}$ our conclusion is that $\mathbf{A}^{\mathsf{T}}$ cannot have more linearly independent columns than rows, so that $\mathbf{A}$ cannot have more linearly independent rows than columns. Together, the number of linearly independent columns of $\mathbf{A}$ must be r, the rank of $\mathbf{A}$. This completes the proof. ■

EXAMPLE 4 Illustration of Theorem 3

The matrix in (2) has rank 2. From Example 3 we see that the first two row vectors are linearly independent and by "working backward" we can verify that Row 3 $= 6$ Row 1 $- \frac{1}{2}$ Row 2. Similarly, the first two columns are linearly independent, and by reducing the last matrix in Example 3 by columns we find that

$$\text{Column 3} = \tfrac{2}{3}\,\text{Column 1} + \tfrac{2}{3}\,\text{Column 2} \qquad \text{and} \qquad \text{Column 4} = \tfrac{2}{3}\,\text{Column 1} + \tfrac{29}{21}\,\text{Column 2}.$$ ■

Combining Theorems 2 and 3 we obtain

THEOREM 4 Linear Dependence of Vectors

Consider p vectors each having n components. If $n < p$, then these vectors *are linearly dependent.*

PROOF The matrix $\mathbf{A}$ with those p vectors as row vectors has p rows and $n < p$ columns; hence by Theorem 3 it has rank $\mathbf{A} \leqq n < p$, which implies linear dependence by Theorem 2. ■

Vector Space

The following related concepts are of general interest in linear algebra. In the present context they provide a clarification of essential properties of matrices and their role in connection with linear systems.

www.EngineeringEBooksPdf.com

Consider a nonempty set V of vectors where each vector has the same number of components. If, for any two vectors **a** and **b** in V, we have that all their linear combinations $\alpha\mathbf{a} + \beta\mathbf{b}$ (α, β any real numbers) are also elements of V, and if, furthermore, **a** and **b** satisfy the laws (3a), (3c), (3d), and (4) in Sec. 7.1, as well as any vectors **a**, **b**, **c** in V satisfy (3b) then V is a vector space. Note that here we wrote laws (3) and (4) of Sec. 7.1 in lowercase letters **a**, **b**, **c**, which is our notation for vectors. More on vector spaces in Sec. 7.9.

The maximum number of linearly independent vectors in V is called the **dimension** of V and is denoted by dim V. Here we assume the dimension to be finite; infinite dimension will be defined in Sec. 7.9.

A linearly independent set in V consisting of a maximum possible number of vectors in V is called a **basis** for V. In other words, any largest possible set of independent vectors in V forms basis for V. That means, if we add one or more vector to that set, the set will be linearly dependent. (See also the beginning of Sec. 7.4 on linear independence and dependence of vectors.) Thus, the number of vectors of a basis for V equals dim V.

The set of all linear combinations of given vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(p)}$ with the same number of components is called the **span** of these vectors. Obviously, a span is a vector space. If in addition, the given vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(p)}$ are linearly independent, then they form a basis for that vector space.

This then leads to another equivalent definition of basis. A set of vectors is a **basis** for a vector space V if (1) the vectors in the set are linearly independent, and if (2) any vector in V can be expressed as a linear combination of the vectors in the set. If (2) holds, we also say that the set of vectors **spans** the vector space V.

By a **subspace** of a vector space V we mean a nonempty subset of V (including V itself) that forms a vector space with respect to the two algebraic operations (addition and scalar multiplication) defined for the vectors of V.

EXAMPLE 5 **Vector Space, Dimension, Basis**

The span of the three vectors in Example 1 is a vector space of dimension 2. A basis of this vector space consists of any two of those three vectors, for instance, $\mathbf{a}_{(1)}, \mathbf{a}_{(2)}$, or $\mathbf{a}_{(1)}, \mathbf{a}_{(3)}$, etc. ■

We further note the simple

THEOREM 5 **Vector Space R^n**

The vector space R^n consisting of all vectors with n components (n real numbers) has dimension n.

PROOF A basis of n vectors is $\mathbf{a}_{(1)} = [1 \quad 0 \quad \cdots \quad 0]$, $\mathbf{a}_{(2)} = [0 \quad 1 \quad 0 \quad \cdots \quad 0]$, $\cdots$, $\mathbf{a}_{(n)} = [0 \quad \cdots \quad 0 \quad 1]$. ■

For a matrix **A**, we call the span of the row vectors the **row space** of **A**. Similarly, the span of the column vectors of **A** is called the **column space** of **A**.

Now, Theorem 3 shows that a matrix **A** has as many linearly independent rows as columns. By the definition of dimension, their number is the dimension of the row space or the column space of **A**. This proves

THEOREM 6 **Row Space and Column Space**

The row space and the column space of a matrix **A** *have the same dimension, equal to rank* **A**.

www.EngineeringEBooksPdf.com

Finally, for a given matrix **A** the solution set of the homogeneous system $\mathbf{Ax} = 0$ is a vector space, called the **null space** of **A**, and its dimension is called the **nullity** of **A**. In the next section we motivate and prove the basic relation

$$\text{rank } \mathbf{A} + \text{nullity } \mathbf{A} = \text{Number of columns of } \mathbf{A}. \tag{6}$$

PROBLEM SET 7.4

1–10 RANK, ROW SPACE, COLUMN SPACE

Find the rank. Find a basis for the row space. Find a basis for the column space. *Hint.* Row-reduce the matrix and its transpose. (You may omit obvious factors from the vectors of these bases.)

1. $\begin{bmatrix} 4 & -2 & 6 \\ -2 & 1 & -3 \end{bmatrix}$

2. $\begin{bmatrix} a & b \\ b & a \end{bmatrix}$

3. $\begin{bmatrix} 0 & 3 & 5 \\ 3 & 5 & 0 \\ 5 & 0 & 10 \end{bmatrix}$

4. $\begin{bmatrix} 6 & -4 & 0 \\ -4 & 0 & 2 \\ 0 & 2 & 6 \end{bmatrix}$

5. $\begin{bmatrix} 0.2 & -0.1 & 0.4 \\ 0 & 1.1 & -0.3 \\ 0.1 & 0 & -2.1 \end{bmatrix}$

6. $\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & -4 \\ 0 & 4 & 0 \end{bmatrix}$

7. $\begin{bmatrix} 8 & 0 & 4 & 0 \\ 0 & 2 & 0 & 4 \\ 4 & 0 & 2 & 0 \end{bmatrix}$

8. $\begin{bmatrix} 2 & 4 & 8 & 16 \\ 16 & 8 & 4 & 2 \\ 4 & 8 & 16 & 2 \\ 2 & 16 & 8 & 4 \end{bmatrix}$

9. $\begin{bmatrix} 9 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

10. $\begin{bmatrix} 5 & -2 & 1 & 0 \\ -2 & 0 & -4 & 1 \\ 1 & -4 & -11 & 2 \\ 0 & 1 & 2 & 0 \end{bmatrix}$

11. CAS Experiment. Rank. (a) Show experimentally that the $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ with $a_{jk} = j + k - 1$ has rank 2 for any n. (Problem 20 shows $n = 4$.) Try to prove it.

(b) Do the same when $a_{jk} = j + k + c$, where c is any positive integer.

(c) What is rank **A** if $a_{jk} = 2^{j+k-2}$? Try to find other large matrices of low rank independent of n.

12–16 GENERAL PROPERTIES OF RANK

Show the following:

12. rank $\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$ = rank **AB**. (Note the order!)

13. rank **A** = rank **B** does *not* imply rank $\mathbf{A}^2$ = rank $\mathbf{B}^2$. (Give a counterexample.)

14. If **A** is not square, either the row vectors or the column vectors of **A** are linearly dependent.

15. If the row vectors of a square matrix are linearly independent, so are the column vectors, and conversely.

16. Give examples showing that the rank of a product of matrices cannot exceed the rank of either factor.

17–25 LINEAR INDEPENDENCE

Are the following sets of vectors linearly independent? Show the details of your work.

17. $[3 \quad 4 \quad 0 \quad 2]$, $[2 \quad -1 \quad 3 \quad 7]$, $[1 \quad 16 \quad -12 \quad -22]$

18. $[1 \quad \frac{1}{2} \quad \frac{1}{3} \quad \frac{1}{4}]$, $[\frac{1}{2} \quad \frac{1}{3} \quad \frac{1}{4} \quad \frac{1}{5}]$, $[\frac{1}{3} \quad \frac{1}{4} \quad \frac{1}{5} \quad \frac{1}{6}]$, $[\frac{1}{4} \quad \frac{1}{5} \quad \frac{1}{6} \quad \frac{1}{7}]$

19. $[0 \quad 1 \quad 1]$, $[1 \quad 1 \quad 1]$, $[0 \quad 0 \quad 1]$

20. $[1 \quad 2 \quad 3 \quad 4]$, $[2 \quad 3 \quad 4 \quad 5]$, $[3 \quad 4 \quad 5 \quad 6]$, $[4 \quad 5 \quad 6 \quad 7]$

21. $[2 \quad 0 \quad 0 \quad 7]$, $[2 \quad 0 \quad 0 \quad 8]$, $[2 \quad 0 \quad 0 \quad 9]$, $[2 \quad 0 \quad 1 \quad 0]$

22. $[0.4 \quad -0.2 \quad 0.2]$, $[0 \quad 0 \quad 0]$, $[3.0 \quad -0.6 \quad 1.5]$

23. $[9 \quad 8 \quad 7 \quad 6 \quad 5]$, $[9 \quad 7 \quad 5 \quad 3 \quad 1]$

24. $[4 \quad -1 \quad 3]$, $[0 \quad 8 \quad 1]$, $[1 \quad 3 \quad -5]$, $[2 \quad 6 \quad 1]$

25. $[6 \quad 0 \quad -1 \quad 3]$, $[2 \quad 2 \quad 5 \quad 0]$, $[-4 \quad -4 \quad -4 \quad -4]$

26. Linearly independent subset. Beginning with the last of the vectors $[3 \quad 0 \quad 1 \quad 2]$, $[6 \quad 1 \quad 0 \quad 0]$, $[12 \quad 1 \quad 2 \quad 4]$, $[6 \quad 0 \quad 2 \quad 4]$, and $[9 \quad 0 \quad 1 \quad 2]$, omit one after another until you get a linearly independent set.

www.EngineeringEBooksPdf.com

27–35 VECTOR SPACE

Is the given set of vectors a vector space? Give reasons. If your answer is yes, determine the dimension and find a basis. ($v_1, v_2, \cdots$ denote components.)

27. All vectors in R^3 with $v_1 - v_2 + 2v_3 = 0$

28. All vectors in R^3 with $3v_2 + v_3 = k$

29. All vectors in R^2 with $v_1 \geqq v_2$

30. All vectors in R^n with the first $n - 2$ components zero

31. All vectors in R^5 with positive components

32. All vectors in R^3 with $3v_1 - 2v_2 + v_3 = 0$, $4v_1 + 5v_2 = 0$

33. All vectors in R^3 with $3v_1 - v_3 = 0$, $2v_1 + 3v_2 - 4v_3 = 0$

34. All vectors in R^n with $|v_j| = 1$ for $j = 1, \cdots, n$

35. All vectors in R^4 with $v_1 = 2v_2 = 3v_3 = 4v_4$

7.5 Solutions of Linear Systems: Existence, Uniqueness

Rank, as just defined, gives complete information about existence, uniqueness, and general structure of the solution set of linear systems as follows.

A linear system of equations in n unknowns has a unique solution if the coefficient matrix and the augmented matrix have the same rank n, and infinitely many solutions if that common rank is less than n. The system has no solution if those two matrices have different rank.

To state this precisely and prove it, we shall use the generally important concept of a **submatrix** of **A**. By this we mean any matrix obtained from **A** by omitting some rows or columns (or both). By definition this includes **A** itself (as the matrix obtained by omitting no rows or columns); this is practical.

THEOREM 1 Fundamental Theorem for Linear Systems

(a) Existence. *A linear system of m equations in n unknowns* $x_1, \cdots, x_n$

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{1}$$

is **consistent**, *that is, has solutions, if and only if the coefficient matrix* **A** *and the augmented matrix* $\tilde{\mathbf{A}}$ *have the same rank. Here,*

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{A}} = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right]$$

(b) Uniqueness. *The system* (1) *has precisely one solution if and only if this common rank r of* **A** *and* $\tilde{\mathbf{A}}$ *equals n.*

www.EngineeringEBooksPdf.com

(c) Infinitely many solutions. *If this common rank r is less than n, the system* (1) *has infinitely many solutions. All of these solutions are obtained by determining r suitable unknowns* (*whose submatrix of coefficients must have rank r*) *in terms of the remaining $n - r$ unknowns, to which arbitrary values can be assigned.* (See Example 3 in Sec. 7.3.)

(d) Gauss elimination (Sec. 7.3). *If solutions exist, they can all be obtained by the Gauss elimination.* (This method will automatically reveal whether or not solutions exist; see Sec. 7.3.)

PROOF **(a)** We can write the system (1) in vector form $\mathbf{Ax} = \mathbf{b}$ or in terms of column vectors $\mathbf{c}_{(1)}, \cdots, \mathbf{c}_{(n)}$ of $\mathbf{A}$:

$$\mathbf{c}_{(1)}x_1 + \mathbf{c}_{(2)}x_2 + \cdots + \mathbf{c}_{(n)}x_n = \mathbf{b}. \tag{2}$$

$\widetilde{\mathbf{A}}$ is obtained by augmenting $\mathbf{A}$ by a single column $\mathbf{b}$. Hence, by Theorem 3 in Sec. 7.4, rank $\widetilde{\mathbf{A}}$ equals rank $\mathbf{A}$ or rank $\mathbf{A} + 1$. Now if (1) has a solution $\mathbf{x}$, then (2) shows that $\mathbf{b}$ must be a linear combination of those column vectors, so that $\widetilde{\mathbf{A}}$ and $\mathbf{A}$ have the same maximum number of linearly independent column vectors and thus the same rank.

Conversely, if rank $\widetilde{\mathbf{A}} =$ rank $\mathbf{A}$, then $\mathbf{b}$ must be a linear combination of the column vectors of $\mathbf{A}$, say,

$$\mathbf{b} = \alpha_1\mathbf{c}_{(1)} + \cdots + \alpha_n\mathbf{c}_{(n)} \tag{2*}$$

since otherwise rank $\widetilde{\mathbf{A}} =$ rank $\mathbf{A} + 1$. But (2*) means that (1) has a solution, namely, $x_1 = \alpha_1, \cdots, x_n = \alpha_n$, as can be seen by comparing (2*) and (2).

(b) If rank $\mathbf{A} = n$, the n column vectors in (2) are linearly independent by Theorem 3 in Sec. 7.4. We claim that then the representation (2) of $\mathbf{b}$ is unique because otherwise

$$\mathbf{c}_{(1)}x_1 + \cdots + \mathbf{c}_{(n)}x_n = \mathbf{c}_{(1)}\widetilde{x}_1 + \cdots + \mathbf{c}_{(n)}\widetilde{x}_n.$$

This would imply (take all terms to the left, with a minus sign)

$$(x_1 - \widetilde{x}_1)\mathbf{c}_{(1)} + \cdots + (x_n - \widetilde{x}_n)\mathbf{c}_{(n)} = \mathbf{0}$$

and $x_1 - \widetilde{x}_1 = 0, \cdots, x_n - \widetilde{x}_n = 0$ by linear independence. But this means that the scalars $x_1, \cdots, x_n$ in (2) are uniquely determined, that is, the solution of (1) is unique.

(c) If rank $\mathbf{A} =$ rank $\widetilde{\mathbf{A}} = r < n$, then by Theorem 3 in Sec. 7.4 there is a linearly independent set K of r column vectors of $\mathbf{A}$ such that the other $n - r$ column vectors of $\mathbf{A}$ are linear combinations of those vectors. We renumber the columns and unknowns, denoting the renumbered quantities by $\hat{}$, so that $\{\hat{\mathbf{c}}_{(1)}, \cdots, \hat{\mathbf{c}}_{(r)}\}$ is that linearly independent set K. Then (2) becomes

$$\hat{\mathbf{c}}_{(1)}\hat{x}_1 + \cdots + \hat{\mathbf{c}}_{(r)}\hat{x}_r + \hat{\mathbf{c}}_{(r+1)}\hat{x}_{r+1} + \cdots + \hat{\mathbf{c}}_{(n)}\hat{x}_n = \mathbf{b},$$

$\hat{\mathbf{c}}_{(r+1)}, \cdots, \hat{\mathbf{c}}_{(n)}$ are linear combinations of the vectors of K, and so are the vectors $\hat{x}_{r+1}\hat{\mathbf{c}}_{(r+1)}, \cdots, \hat{x}_n\hat{\mathbf{c}}_{(n)}$. Expressing these vectors in terms of the vectors of K and collecting terms, we can thus write the system in the form

$$\hat{\mathbf{c}}_{(1)}y_1 + \cdots + \hat{\mathbf{c}}_{(r)}y_r = \mathbf{b} \tag{3}$$

www.EngineeringEBooksPdf.com

with $y_j = \hat{x}_j + \beta_j$, where β_j results from the $n - r$ terms $\hat{\mathbf{c}}_{(r+1)}\hat{x}_{r+1}, \cdots, \hat{\mathbf{c}}_{(n)}\hat{x}_n$; here, $j = 1, \cdots, r$. Since the system has a solution, there are $y_1, \cdots, y_r$ satisfying (3). These scalars are unique since K is linearly independent. Choosing $\hat{x}_{r+1}, \cdots, \hat{x}_n$ fixes the β_j and corresponding $\hat{x}_j = y_j - \beta_j$, where $j = 1, \cdots, r$.

(d) This was discussed in Sec. 7.3 and is restated here as a reminder. ■

The theorem is illustrated in Sec. 7.3. In Example 2 there is a unique solution since rank $\widetilde{\mathbf{A}} = \text{rank } \mathbf{A} = n = 3$ (as can be seen from the last matrix in the example). In Example 3 we have rank $\widetilde{\mathbf{A}} = \text{rank } \mathbf{A} = 2 < n = 4$ and can choose x_3 and x_4 arbitrarily. In Example 4 there is no solution because rank $\mathbf{A} = 2 <$ rank $\widetilde{\mathbf{A}} = 3$.

Homogeneous Linear System

Recall from Sec. 7.3 that a linear system (1) is called **homogeneous** if all the b_j's are zero, and **nonhomogeneous** if one or several b_j's are not zero. For the homogeneous system we obtain from the Fundamental Theorem the following results.

THEOREM 2

Homogeneous Linear System

A homogeneous linear system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= 0 \end{aligned} \tag{4}$$

always has the **trivial solution** $x_1 = 0, \cdots, x_n = 0$. *Nontrivial solutions exist if and only if rank* $\mathbf{A} < n$. *If* rank $\mathbf{A} = r < n$, *these solutions, together with* $\mathbf{x} = \mathbf{0}$, *form a vector space* (see Sec. 7.4) *of dimension* $n - r$ called the **solution space** of (4).

In particular, if $\mathbf{x}_{(1)}$ *and* $\mathbf{x}_{(2)}$ *are solution vectors of* (4), *then* $\mathbf{x} = c_1\mathbf{x}_{(1)} + c_2\mathbf{x}_{(2)}$ *with any scalars* c_1 *and* c_2 *is a solution vector of* (4). (This ***does not hold*** for nonhomogeneous systems. Also, the term *solution space is used for homogeneous* systems only.)

PROOF The first proposition can be seen directly from the system. It agrees with the fact that $\mathbf{b} = \mathbf{0}$ implies that rank $\widetilde{\mathbf{A}} =$ rank $\mathbf{A}$, so that a homogeneous system is always ***consistent***. If rank $\mathbf{A} = n$, the trivial solution is the unique solution according to (b) in Theorem 1. If rank $\mathbf{A} < n$, there are nontrivial solutions according to (c) in Theorem 1. The solutions form a vector space because if $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ are any of them, then $\mathbf{A}\mathbf{x}_{(1)} = \mathbf{0}$, $\mathbf{A}\mathbf{x}_{(2)} = \mathbf{0}$, and this implies $\mathbf{A}(\mathbf{x}_{(1)} + \mathbf{x}_{(2)}) = \mathbf{A}\mathbf{x}_{(1)} + \mathbf{A}\mathbf{x}_{(2)} = \mathbf{0}$ as well as $\mathbf{A}(c\mathbf{x}_{(1)}) = c\mathbf{A}\mathbf{x}_{(1)} = \mathbf{0}$, where c is arbitrary. If rank $\mathbf{A} = r < n$, Theorem 1 (c) implies that we can choose $n - r$ suitable unknowns, call them $x_{r+1}, \cdots, x_n$, in an arbitrary fashion, and every solution is obtained in this way. Hence a basis for the solution space, briefly called a **basis of solutions** of (4), is $\mathbf{y}_{(1)}, \cdots, \mathbf{y}_{(n-r)}$, where the basis vector $\mathbf{y}_{(j)}$ is obtained by choosing $x_{r+j} = 1$ and the other $x_{r+1}, \cdots, x_n$ zero; the corresponding first r components of this solution vector are then determined. Thus the solution space of (4) has dimension $n - r$. This proves Theorem 2. ■

www.EngineeringEBooksPdf.com

The solution space of (4) is also called the **null space** of $\mathbf{A}$ because $\mathbf{Ax} = \mathbf{0}$ for every $\mathbf{x}$ in the solution space of (4). Its dimension is called the **nullity** of $\mathbf{A}$. Hence Theorem 2 states that

$$\text{rank } \mathbf{A} + \text{nullity } \mathbf{A} = n \tag{5}$$

where n is the number of unknowns (number of columns of $\mathbf{A}$).

Furthermore, by the definition of rank we have rank $\mathbf{A} \leqq m$ in (4). Hence if $m < n$, then rank $\mathbf{A} < n$. By Theorem 2 this gives the practically important

THEOREM 3 **Homogeneous Linear System with Fewer Equations Than Unknowns**

A homogeneous linear system with fewer equations than unknowns always has nontrivial solutions.

Nonhomogeneous Linear Systems

The characterization of all solutions of the linear system (1) is now quite simple, as follows.

THEOREM 4 **Nonhomogeneous Linear System**

If a nonhomogeneous linear system (1) *is consistent, then all of its solutions are obtained as*

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{x}_h \tag{6}$$

where $\mathbf{x}_0$ *is any* (fixed) *solution of* (1) *and* $\mathbf{x}_h$ *runs through all the solutions of the corresponding homogeneous system* (4).

PROOF The difference $\mathbf{x}_h = \mathbf{x} - \mathbf{x}_0$ of any two solutions of (1) is a solution of (4) because $\mathbf{Ax}_h = \mathbf{A}(\mathbf{x} - \mathbf{x}_0) = \mathbf{Ax} - \mathbf{Ax}_0 = \mathbf{b} - \mathbf{b} = \mathbf{0}$. Since $\mathbf{x}$ is any solution of (1), we get all the solutions of (1) if in (6) we take any solution $\mathbf{x}_0$ of (1) and let $\mathbf{x}_h$ vary throughout the solution space of (4). ■

This covers a main part of our discussion of characterizing the solutions of systems of linear equations. Our next main topic is determinants and their role in linear equations.

7.6 For Reference: Second- and Third-Order Determinants

We created this section as a quick general reference section on second- and third-order determinants. It is completely independent of the theory in Sec. 7.7 and suffices as a reference for many of our examples and problems. Since this section is for reference, ***go on to the next section, consulting this material only when needed***.

A **determinant of second order** is denoted and defined by

$$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{1}$$

So here we have ***bars*** (whereas a matrix has ***brackets***).

www.EngineeringEBooksPdf.com

Cramer's rule for solving linear systems of two equations in two unknowns

$$(2)\qquad \begin{aligned} &(a)\quad a_{11}x_1 + a_{12}x_2 = b_1 \\ &(b)\quad a_{21}x_1 + a_{22}x_2 = b_2 \end{aligned}$$

is

$$(3)\qquad \begin{aligned} x_1 &= \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{D} = \frac{b_1a_{22} - a_{12}b_2}{D}, \\ x_2 &= \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{D} = \frac{a_{11}b_2 - b_1a_{21}}{D} \end{aligned}$$

with D as in (1), provided

$$D \neq 0.$$

The value $D = 0$ appears for homogeneous systems with nontrivial solutions.

PROOF We prove (3). To eliminate x_2 multiply (2a) by a_{22} and (2b) by $-a_{12}$ and add,

$$(a_{11}a_{22} - a_{12}a_{21})x_1 = b_1a_{22} - a_{12}b_2.$$

Similarly, to eliminate x_1 multiply (2a) by $-a_{21}$ and (2b) by a_{11} and add,

$$(a_{11}a_{22} - a_{12}a_{21})x_2 = a_{11}b_2 - b_1a_{21}.$$

Assuming that $D = a_{11}a_{22} - a_{12}a_{21} \neq 0$, dividing, and writing the right sides of these two equations as determinants, we obtain (3). ■

EXAMPLE 1 Cramer's Rule for Two Equations

If $\begin{aligned} 4x_1 + 3x_2 &= 12 \\ 2x_1 + 5x_2 &= -8 \end{aligned}$ then $x_1 = \dfrac{\begin{vmatrix} 12 & 3 \\ -8 & 5 \end{vmatrix}}{\begin{vmatrix} 4 & 3 \\ 2 & 5 \end{vmatrix}} = \dfrac{84}{14} = 6, \quad x_2 = \dfrac{\begin{vmatrix} 4 & 12 \\ 2 & -8 \end{vmatrix}}{\begin{vmatrix} 4 & 3 \\ 2 & 5 \end{vmatrix}} = \dfrac{-56}{14} = -4.$ ■

Third-Order Determinants

A **determinant of third order** can be defined by

$$(4)\qquad D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21}\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31}\begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix}.$$

www.EngineeringEBooksPdf.com

Note the following. The signs on the right are $+ - +$. Each of the three terms on the right is an entry in the first column of D times its **minor**, that is, the second-order determinant obtained from D by deleting the row and column of that entry; thus, for a_{11} delete the first row and first column, and so on.

If we write out the minors in (4), we obtain

$$D = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{21}a_{13}a_{32} - a_{21}a_{12}a_{33} + a_{31}a_{12}a_{23} - a_{31}a_{13}a_{22}. \tag{4*}$$

Cramer's Rule for Linear Systems of Three Equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \tag{5}$$

is

$$x_1 = \frac{D_1}{D}, \quad x_2 = \frac{D_2}{D}, \quad x_3 = \frac{D_3}{D} \qquad (D \neq 0) \tag{6}$$

with the *determinant D of the system* given by (4) and

$$D_1 = \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}, \quad D_2 = \begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix}, \quad D_3 = \begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}.$$

Note that D_1, D_2, D_3 are obtained by replacing Columns 1, 2, 3, respectively, by the column of the right sides of (5).

Cramer's rule (6) can be derived by eliminations similar to those for (3), but it also follows from the general case (Theorem 4) in the next section.

7.7 Determinants. Cramer's Rule

Determinants were originally introduced for solving linear systems. Although ***impractical in computations***, they have important engineering applications in eigenvalue problems (Sec. 8.1), differential equations, vector algebra (Sec. 9.3), and in other areas. They can be introduced in several equivalent ways. Our definition is particularly for dealing with linear systems.

A **determinant of order** n is a scalar associated with an $n \times n$ (hence ***square*!**) matrix $\mathbf{A} = [a_{jk}]$, and is denoted by

$$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}. \tag{1}$$

www.EngineeringEBooksPdf.com

For $n = 1$, this determinant is defined by

$$D = a_{11}. \tag{2}$$

For $n \geqq 2$ by

$$D = a_{j1}C_{j1} + a_{j2}C_{j2} + \cdots + a_{jn}C_{jn} \qquad (j = 1, 2, \cdots, \text{or } n) \tag{3a}$$

or

$$D = a_{1k}C_{1k} + a_{2k}C_{2k} + \cdots + a_{nk}C_{nk} \quad (k = 1, 2, \cdots, \text{or } n). \tag{3b}$$

Here,

$$C_{jk} = (-1)^{j+k}M_{jk}$$

and M_{jk} is a determinant of order $n - 1$, namely, the determinant of the submatrix of $\mathbf{A}$ obtained from $\mathbf{A}$ by omitting the row and column of the entry a_{jk}, that is, the jth row and the kth column.

In this way, D is defined in terms of n determinants of order $n - 1$, each of which is, in turn, defined in terms of $n - 1$ determinants of order $n - 2$, and so on—until we finally arrive at second-order determinants, in which those submatrices consist of single entries whose determinant is defined to be the entry itself.

From the definition it follows that *we may **expand** D by any row or column,* that is, choose in (3) the entries in any row or column, similarly when expanding the C_{jk}'s in (3), and so on.

This definition is unambiguous, that is, it yields the same value for D no matter which columns or rows we choose in expanding. A proof is given in App. 4.

Terms used in connection with determinants are taken from matrices. In D we have n^2 **entries** a_{jk}, also n **rows** and n **columns**, and a **main diagonal** on which $a_{11}, a_{22}, \cdots, a_{nn}$ stand. Two terms are new:

M_{jk} is called the **minor** *of* a_{jk} *in D,* and C_{jk} the **cofactor** *of* a_{jk} *in D.*

For later use we note that (3) may also be written in terms of minors

$$D = \sum_{k=1}^{n} (-1)^{j+k} a_{jk} M_{jk} \qquad (j = 1, 2, \cdots, \text{or } n) \tag{4a}$$

$$D = \sum_{j=1}^{n} (-1)^{j+k} a_{jk} M_{jk} \qquad (k = 1, 2, \cdots, \text{or } n). \tag{4b}$$

EXAMPLE 1 Minors and Cofactors of a Third-Order Determinant

In (4) of the previous section the minors and cofactors of the entries in the first column can be seen directly. For the entries in the second row the minors are

$$M_{21} = \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix}, \qquad M_{22} = \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}, \qquad M_{23} = \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix}$$

and the cofactors are $C_{21} = -M_{21}$, $C_{22} = +M_{22}$, and $C_{23} = -M_{23}$. Similarly for the third row—write these down yourself. And verify that the signs in C_{jk} form a **checkerboard pattern**

$$\begin{matrix} + & - & + \\ - & + & - \\ + & - & + \end{matrix}$$

www.EngineeringEBooksPdf.com

EXAMPLE 2 Expansions of a Third-Order Determinant

$$D = \begin{vmatrix} 1 & 3 & 0 \\ 2 & 6 & 4 \\ -1 & 0 & 2 \end{vmatrix} = 1\begin{vmatrix} 6 & 4 \\ 0 & 2 \end{vmatrix} - 3\begin{vmatrix} 2 & 4 \\ -1 & 2 \end{vmatrix} + 0\begin{vmatrix} 2 & 6 \\ -1 & 0 \end{vmatrix}$$

$$= 1(12 - 0) - 3(4 + 4) + 0(0 + 6) = -12.$$

This is the expansion by the first row. The expansion by the third column is

$$D = 0\begin{vmatrix} 2 & 6 \\ -1 & 0 \end{vmatrix} - 4\begin{vmatrix} 1 & 3 \\ -1 & 0 \end{vmatrix} + 2\begin{vmatrix} 1 & 3 \\ 2 & 6 \end{vmatrix} = 0 - 12 + 0 = -12.$$

Verify that the other four expansions also give the value -12. ■

EXAMPLE 3 Determinant of a Triangular Matrix

$$\begin{vmatrix} -3 & 0 & 0 \\ 6 & 4 & 0 \\ -1 & 2 & 5 \end{vmatrix} = -3\begin{vmatrix} 4 & 0 \\ 2 & 5 \end{vmatrix} = -3 \cdot 4 \cdot 5 = -60.$$

Inspired by this, can you formulate a little theorem on determinants of triangular matrices? Of diagonal matrices? ■

General Properties of Determinants

There is an attractive way of finding determinants (1) that consists of applying elementary row operations to (1). By doing so we obtain an "upper triangular" determinant (see Sec. 7.1, for definition with "matrix" replaced by "determinant") whose value is then very easy to compute, being just the product of its diagonal entries. This approach is ***similar*** (***but not the same!***) to what we did to matrices in Sec. 7.3. *In particular, be aware that interchanging two rows in a determinant introduces a multiplicative factor of* -1 *to the value of the determinant*! Details are as follows.

THEOREM 1 Behavior of an *n*th-Order Determinant under Elementary Row Operations

(a) *Interchange of two rows multiplies the value of the determinant by* -1.

(b) *Addition of a multiple of a row to another row does not alter the value of the determinant.*

(c) *Multiplication of a row by a nonzero constant c multiplies the value of the determinant by c.* (This holds also when $c = 0$, but no longer gives an elementary row operation.)

PROOF **(a)** By induction. The statement holds for $n = 2$ because

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc, \qquad \text{but} \qquad \begin{vmatrix} c & d \\ a & b \end{vmatrix} = bc - ad.$$

www.EngineeringEBooksPdf.com

We now make the induction hypothesis that (a) holds for determinants of order $n - 1 \geqq 2$ and show that it then holds for determinants of order n. Let D be of order n. Let E be obtained from D by the interchange of two rows. Expand D and E by a row that is ***not*** one of those interchanged, call it the jth row. Then by (4a),

$$D = \sum_{k=1}^{n} (-1)^{j+k} a_{jk} M_{jk}, \qquad E = \sum_{k=1}^{n} (-1)^{j+k} a_{jk} N_{jk} \tag{5}$$

where N_{jk} is obtained from the minor M_{jk} of a_{jk} in D by the interchange of those two rows which have been interchanged in D (and which N_{jk} must both contain because we expand by another row!). Now these minors are of order $n - 1$. Hence the induction hypothesis applies and gives $N_{jk} = -M_{jk}$. Thus $E = -D$ by (5).

(b) Add c times Row i to Row j. Let $\tilde{D}$ be the new determinant. Its entries in Row j are $a_{jk} + ca_{ik}$. If we expand $\tilde{D}$ by this Row j, we see that we can write it as $\tilde{D} = D_1 + cD_2$, where $D_1 = D$ has in Row j the a_{jk}, whereas D_2 has in that Row j the a_{jk} from the addition. Hence D_2 has a_{jk} in both Row i and Row j. Interchanging these two rows gives D_2 back, but on the other hand it gives $-D_2$ by (a). Together $D_2 = -D_2 = 0$, so that $\tilde{D} = D_1 = D$.

(c) Expand the determinant by the row that has been multiplied.

CAUTION! $\det(c\mathbf{A}) = c^n \det \mathbf{A}$ (not $c \det \mathbf{A}$). Explain why. ■

EXAMPLE 4 Evaluation of Determinants by Reduction to Triangular Form

Because of Theorem 1 we may evaluate determinants by reduction to triangular form, as in the Gauss elimination for a matrix. For instance (with the blue explanations always referring to the ***preceding determinant***)

$$D = \begin{vmatrix} 2 & 0 & -4 & 6 \\ 4 & 5 & 1 & 0 \\ 0 & 2 & 6 & -1 \\ -3 & 8 & 9 & 1 \end{vmatrix}$$

$$= \begin{vmatrix} 2 & 0 & -4 & 6 \\ 0 & 5 & 9 & -12 \\ 0 & 2 & 6 & -1 \\ 0 & 8 & 3 & 10 \end{vmatrix} \begin{array}{l} \\ \text{Row 2} - 2\,\text{Row 1} \\ \\ \text{Row 4} + 1.5\,\text{Row 1} \end{array}$$

$$= \begin{vmatrix} 2 & 0 & -4 & 6 \\ 0 & 5 & 9 & -12 \\ 0 & 0 & 2.4 & 3.8 \\ 0 & 0 & -11.4 & 29.2 \end{vmatrix} \begin{array}{l} \\ \\ \text{Row 3} - 0.4\,\text{Row 2} \\ \text{Row 4} - 1.6\,\text{Row 2} \end{array}$$

$$= \begin{vmatrix} 2 & 0 & -4 & 6 \\ 0 & 5 & 9 & -12 \\ 0 & 0 & 2.4 & 3.8 \\ 0 & 0 & -0 & 47.25 \end{vmatrix} \begin{array}{l} \\ \\ \\ \text{Row 4} + 4.75\,\text{Row 3} \end{array}$$

$$= 2 \cdot 5 \cdot 2.4 \cdot 47.25 = 1134.$$ ■

www.EngineeringEBooksPdf.com

THEOREM 2 **Further Properties of *n*th-Order Determinants**

(a)–(c) *in Theorem* 1 *hold also for columns.*

(d) ***Transposition*** *leaves the value of a determinant unaltered.*

(e) ***A zero row or column*** *renders the value of a determinant zero.*

(f) ***Proportional rows or columns*** *render the value of a determinant zero. In particular, a determinant with two identical rows or columns has the value zero.*

PROOF **(a)–(e)** follow directly from the fact that a determinant can be expanded by any row column. In **(d)**, transposition is defined as for matrices, that is, the jth row becomes the jth column of the transpose.

(f) If Row $j = c$ times Row i, then $D = cD_1$, where D_1 has Row j = Row i. Hence an interchange of these rows reproduces D_1, but it also gives $-D_1$ by Theorem 1(a). Hence $D_1 = 0$ and $D = cD_1 = 0$. Similarly for columns. ■

It is quite remarkable that the important concept of the rank of a matrix $\mathbf{A}$, which is the maximum number of linearly independent row or column vectors of $\mathbf{A}$ (see Sec. 7.4), can be related to determinants. Here we may assume that rank $\mathbf{A} > 0$ because the only matrices with rank 0 are the zero matrices (see Sec. 7.4).

THEOREM 3 **Rank in Terms of Determinants**

Consider an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$:

(1) $\mathbf{A}$ *has rank $r \geqq 1$ if and only if* $\mathbf{A}$ *has an $r \times r$ submatrix with a nonzero determinant.*

(2) *The determinant of any square submatrix with more than r rows, contained in* $\mathbf{A}$ *(if such a matrix exists!) has a value equal to zero.*

Furthermore, if $m = n$, we have:

(3) *An $n \times n$ square matrix* $\mathbf{A}$ *has rank n if and only if*

$$\det \mathbf{A} \neq 0.$$

PROOF The key idea is that elementary row operations (Sec. 7.3) alter neither rank (by Theorem 1 in Sec. 7.4) nor the property of a determinant being nonzero (by Theorem 1 in this section). The echelon form $\hat{\mathbf{A}}$ of $\mathbf{A}$ (see Sec. 7.3) has r nonzero row vectors (which are the first r row vectors) if and only if rank $\mathbf{A} = r$. Without loss of generality, we can assume that $r \geqq 1$. Let $\hat{\mathbf{R}}$ be the $r \times r$ submatrix in the left upper corner of $\hat{\mathbf{A}}$ (so that the entries of $\hat{\mathbf{R}}$ are in both the first r rows and r columns of $\hat{\mathbf{A}}$). Now $\hat{\mathbf{R}}$ is triangular, with all diagonal entries r_{jj} nonzero. Thus, $\det \hat{\mathbf{R}} = r_{11} \cdots r_{rr} \neq 0$. Also $\det \mathbf{R} \neq 0$ for the corresponding $r \times r$ submatrix $\mathbf{R}$ of $\mathbf{A}$ because $\hat{\mathbf{R}}$ results from $\mathbf{R}$ by elementary row operations. This proves part (1).

Similarly, $\det \mathbf{S} = 0$ for any square submatrix $\mathbf{S}$ of $r + 1$ or more rows perhaps contained in $\mathbf{A}$ because the corresponding submatrix $\hat{\mathbf{S}}$ of $\hat{\mathbf{A}}$ must contain a row of zeros (otherwise we would have rank $\mathbf{A} \geqq r + 1$), so that $\det \hat{\mathbf{S}} = 0$ by Theorem 2. This proves part (2). Furthermore, we have proven the theorem for an $m \times n$ matrix.

www.EngineeringEBooksPdf.com

For an $n \times n$ square matrix $\mathbf{A}$ we proceed as follows. To prove (3), we apply part (1) (already proven!). This gives us that rank $\mathbf{A} = n \geqq 1$ if and only if $\mathbf{A}$ contains an $n \times n$ submatrix with nonzero determinant. But the only such submatrix contained in our square matrix $\mathbf{A}$, is $\mathbf{A}$ itself, hence $\det \mathbf{A} \neq 0$. This proves part (3). ■

Cramer's Rule

Theorem 3 opens the way to the classical solution formula for linear systems known as **Cramer's rule**,[2] which gives solutions as quotients of determinants. ***Cramer's rule is not practical in computations*** for which the methods in Secs. 7.3 and 20.1–20.3 are suitable. However, Cramer's rule is of ***theoretical interest*** in differential equations (Secs. 2.10 and 3.3) and in other theoretical work that has engineering applications.

THEOREM 4

Cramer's Theorem (Solution of Linear Systems by Determinants)

(a) *If a linear system of n equations in the same number of unknowns* $x_1, \cdots, x_n$

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned} \tag{6}$$

has a nonzero coefficient determinant $D = \det \mathbf{A}$, *the system has precisely one solution. This solution is given by the formulas*

$$x_1 = \frac{D_1}{D}, \quad x_2 = \frac{D_2}{D}, \cdots, \quad x_n = \frac{D_n}{D} \tag{7}$$

(Cramer's rule)

where D_k *is the determinant obtained from* D *by replacing in* D *the kth column by the column with the entries* $b_1, \cdots, b_n$.

(b) *Hence if the system* (6) *is* ***homogeneous*** *and* $D \neq 0$, *it has only the trivial solution* $x_1 = 0, x_2 = 0, \cdots, x_n = 0$. *If* $D = 0$, *the homogeneous system also has nontrivial solutions.*

PROOF The augmented matrix $\tilde{\mathbf{A}}$ of the system (6) is of size $n \times (n + 1)$. Hence its rank can be at most n. Now if

$$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} \neq 0, \tag{8}$$

[2]GABRIEL CRAMER (1704–1752), Swiss mathematician.

www.EngineeringEBooksPdf.com

then rank $\mathbf{A} = n$ by Theorem 3. Thus rank $\tilde{\mathbf{A}} = \text{rank}\,\mathbf{A}$. Hence, by the Fundamental Theorem in Sec. 7.5, the system (6) has a unique solution.

Let us now prove (7). Expanding D by its kth column, we obtain

$$D = a_{1k}C_{1k} + a_{2k}C_{2k} + \cdots + a_{nk}C_{nk}, \tag{9}$$

where C_{ik} is the cofactor of entry a_{ik} in D. If we replace the entries in the kth column of D by any other numbers, we obtain a new determinant, say, $\hat{D}$. Clearly, its expansion by the kth column will be of the form (9), with $a_{1k}, \cdots, a_{nk}$ replaced by those new numbers and the cofactors C_{ik} as before. In particular, if we choose as new numbers the entries $a_{1l}, \cdots, a_{nl}$ of the lth column of D (where $l \neq k$), we have a new determinant $\hat{D}$ which has the column $[a_{1l} \quad \cdots \quad a_{nl}]^\mathsf{T}$ twice, once as its lth column, and once as its kth because of the replacement. Hence $\hat{D} = 0$ by Theorem 2(f). If we now expand $\hat{D}$ by the column that has been replaced (the kth column), we thus obtain

$$a_{1l}C_{1k} + a_{2l}C_{2k} + \cdots + a_{nl}C_{nk} = 0 \qquad (l \neq k). \tag{10}$$

We now multiply the first equation in (6) by C_{1k} on both sides, the second by $C_{2k}, \cdots$, the last by C_{nk}, and add the resulting equations. This gives

$$\begin{aligned} C_{1k}(a_{11}x_1 + \cdots + a_{1n}x_n) + \cdots + C_{nk}(a_{n1}x_1 + \cdots + a_{nn}x_n) \\ = b_1C_{1k} + \cdots + b_nC_{nk}. \end{aligned} \tag{11}$$

Collecting terms with the same x_j, we can write the left side as

$$x_1(a_{11}C_{1k} + a_{21}C_{2k} + \cdots + a_{n1}C_{nk}) + \cdots + x_n(a_{1n}C_{1k} + a_{2n}C_{2k} + \cdots + a_{nn}C_{nk}).$$

From this we see that x_k is multiplied by

$$a_{1k}C_{1k} + a_{2k}C_{2k} + \cdots + a_{nk}C_{nk}.$$

Equation (9) shows that this equals D. Similarly, x_1 is multiplied by

$$a_{1l}C_{1k} + a_{2l}C_{2k} + \cdots + a_{nl}C_{nk}.$$

Equation (10) shows that this is zero when $l \neq k$. Accordingly, the left side of (11) equals simply x_kD, so that (11) becomes

$$x_kD = b_1C_{1k} + b_2C_{2k} + \cdots + b_nC_{nk}.$$

Now the right side of this is D_k as defined in the theorem, expanded by its kth column, so that division by D gives (7). This proves Cramer's rule.

If (6) is homogeneous and $D \neq 0$, then each D_k has a column of zeros, so that $D_k = 0$ by Theorem 2(e), and (7) gives the trivial solution.

Finally, if (6) is homogeneous and $D = 0$, then rank $\mathbf{A} < n$ by Theorem 3, so that nontrivial solutions exist by Theorem 2 in Sec. 7.5. ■

EXAMPLE 5 Illustration of Cramer's Rule (Theorem 4)

For $n = 2$, see Example 1 of Sec. 7.6. Also, at the end of that section, we give Cramer's rule for a general linear system of three equations. ■

www.EngineeringEBooksPdf.com

Finally, an important application for Cramer's rule dealing with inverse matrices will be given in the next section.

PROBLEM SET 7.7

1–6 GENERAL PROBLEMS

1. **General Properties of Determinants.** Illustrate each statement in Theorems 1 and 2 with an example of your choice.

2. **Second-Order Determinant.** Expand a general second-order determinant in four possible ways and show that the results agree.

3. **Third-Order Determinant.** Do the task indicated in Theorem 2. Also evaluate D by reduction to triangular form.

4. **Expansion Numerically Impractical.** Show that the computation of an nth-order determinant by expansion involves $n!$ multiplications, which if a multiplication takes 10^{-9} sec would take these times:

n	10	15	20	25
Time	0.004 sec	22 min	77 years	$0.5 \cdot 10^9$ years

5. **Multiplication by Scalar.** Show that $\det(k\mathbf{A}) = k^n \det \mathbf{A}$ (not $k \det \mathbf{A}$). Give an example.

6. **Minors, cofactors.** Complete the list in Example 1.

7–15 EVALUATION OF DETERMINANTS

Showing the details, evaluate:

7. $\begin{vmatrix} \cos\alpha & \sin\alpha \\ \sin\beta & \cos\beta \end{vmatrix}$

8. $\begin{vmatrix} 0.4 & 4.9 \\ 1.5 & -1.3 \end{vmatrix}$

9. $\begin{vmatrix} \cos n\theta & \sin n\theta \\ -\sin n\theta & \cos n\theta \end{vmatrix}$

10. $\begin{vmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{vmatrix}$

11. $\begin{vmatrix} 4 & -1 & 8 \\ 0 & 2 & 3 \\ 0 & 0 & 5 \end{vmatrix}$

12. $\begin{vmatrix} a & b & c \\ c & a & b \\ b & c & a \end{vmatrix}$

13. $\begin{vmatrix} 0 & 4 & -1 & 5 \\ -4 & 0 & 3 & -2 \\ 1 & -3 & 0 & 1 \\ -5 & 2 & -1 & 0 \end{vmatrix}$

14. $\begin{vmatrix} 4 & 7 & 0 & 0 \\ 2 & 8 & 0 & 0 \\ 0 & 0 & 1 & 5 \\ 0 & 0 & -2 & 2 \end{vmatrix}$

15. $\begin{vmatrix} 1 & 2 & 0 & 0 \\ 2 & 4 & 2 & 0 \\ 0 & 2 & 9 & 2 \\ 0 & 0 & 2 & 16 \end{vmatrix}$

16. **CAS EXPERIMENT. Determinant of Zeros and Ones.** Find the value of the determinant of the $n \times n$ matrix A_n with main diagonal entries all 0 and all others 1. Try to find a formula for this. Try to prove it by induction. Interpret A_3 and A_4 as *incidence matrices* (as in Problem Set 7.1 but without the minuses) of a triangle and a tetrahedron, respectively; similarly for an *n-simplex*, having n vertices and $n(n-1)/2$ edges (and spanning R^{n-1}, $n = 5, 6, \cdots$).

17–19 RANK BY DETERMINANTS

Find the rank by Theorem 3 (which is not very practical) and check by row reduction. Show details.

17. $\begin{bmatrix} 4 & 9 \\ -8 & -6 \\ 16 & 12 \end{bmatrix}$

18. $\begin{bmatrix} 0 & 4 & -6 \\ 4 & 0 & 10 \\ -6 & 10 & 0 \end{bmatrix}$

19. $\begin{bmatrix} 1 & 5 & 2 & 2 \\ 1 & 3 & 2 & 6 \\ 4 & 0 & 8 & 48 \end{bmatrix}$

20. **TEAM PROJECT. Geometric Applications: Curves and Surfaces Through Given Points.** The idea is to get an equation from the vanishing of the determinant of a homogeneous linear system as the condition for a nontrivial solution in Cramer's theorem. We explain the trick for obtaining such a system for the case of a line L through two given points P_1: (x_1, y_1) and P_2: (x_2, y_2). The unknown line is $ax + by = -c$, say. We write it as $ax + by + c \cdot 1 = 0$. To get a nontrivial solution a, b, c, the determinant of the "coefficients" $x, y, 1$ must be zero. The system is

$$
\begin{aligned}
ax + by + c \cdot 1 &= 0 \quad (\text{Line } L) \\
ax_1 + by_1 + c \cdot 1 &= 0 \quad (P_1 \text{ on } L) \qquad (12) \\
ax_2 + by_2 + c \cdot 1 &= 0 \quad (P_2 \text{ on } L).
\end{aligned}
$$

www.EngineeringEBooksPdf.com

(a) Line through two points. Derive from $D = 0$ in (12) the familiar formula

$$\frac{x - x_1}{x_1 - x_2} = \frac{y - y_1}{y_1 - y_2}.$$

(b) Plane. Find the analog of (12) for a plane through three given points. Apply it when the points are (1, 1, 1), (3, 2, 6), (5, 0, 5).

(c) Circle. Find a similar formula for a circle in the plane through three given points. Find and sketch the circle through (2, 6), (6, 4), (7, 1).

(d) Sphere. Find the analog of the formula in (c) for a sphere through four given points. Find the sphere through (0, 0, 5), (4, 0, 1), (0, 4, 1), (0, 0, −3) by this formula or by inspection.

(e) General conic section. Find a formula for a general conic section (the vanishing of a determinant of 6th order). Try it out for a quadratic parabola and for a more general conic section of your own choice.

21–25 CRAMER'S RULE

Solve by Cramer's rule. Check by Gauss elimination and back substitution. Show details.

21. $\begin{aligned} 3x - 5y &= 15.5 \\ 6x + 16y &= 5.0 \end{aligned}$

22. $\begin{aligned} 2x - 4y &= -24 \\ 5x + 2y &= 0 \end{aligned}$

23. $\begin{aligned} 3y - 4z &= 16 \\ 2x - 5y + 7z &= -27 \\ -x \qquad - 9z &= 9 \end{aligned}$

24. $\begin{aligned} 3x - 2y + z &= 13 \\ -2x + y + 4z &= 11 \\ x + 4y - 5z &= -31 \end{aligned}$

25. $\begin{aligned} -4w + x + y \qquad &= -10 \\ w - 4x \qquad + z &= 1 \\ w \qquad - 4y + z &= -7 \\ x + y - 4z &= 10 \end{aligned}$

7.8 Inverse of a Matrix. Gauss–Jordan Elimination

***In this section we consider* square *matrices exclusively*.**

The **inverse** of an $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is denoted by $\mathbf{A}^{-1}$ and is an $n \times n$ matrix such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{1}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix (see Sec. 7.2).

If $\mathbf{A}$ has an inverse, then $\mathbf{A}$ is called a **nonsingular matrix**. If $\mathbf{A}$ has no inverse, then $\mathbf{A}$ is called a **singular matrix**.

If $\mathbf{A}$ *has an inverse, the inverse is unique.*

Indeed, if both $\mathbf{B}$ and $\mathbf{C}$ are inverses of $\mathbf{A}$, then $\mathbf{AB} = \mathbf{I}$ and $\mathbf{CA} = \mathbf{I}$, so that we obtain the uniqueness from

$$\mathbf{B} = \mathbf{IB} = (\mathbf{CA})\mathbf{B} = \mathbf{C}(\mathbf{AB}) = \mathbf{CI} = \mathbf{C}.$$

We prove next that $\mathbf{A}$ has an inverse (is nonsingular) if and only if it has maximum possible rank n. The proof will also show that $\mathbf{Ax} = \mathbf{b}$ implies $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ provided $\mathbf{A}^{-1}$ exists, and will thus give a motivation for the inverse as well as a relation to linear systems. (But this will ***not*** give a good method of solving $\mathbf{Ax} = \mathbf{b}$ ***numerically*** because the Gauss elimination in Sec. 7.3 requires fewer computations.)

THEOREM 1

Existence of the Inverse

The inverse $\mathbf{A}^{-1}$ *of an* $n \times n$ *matrix* $\mathbf{A}$ *exists if and only if* rank $\mathbf{A} = n$, *thus* (by Theorem 3, Sec. 7.7) *if and only if* det $\mathbf{A} \neq 0$. *Hence* $\mathbf{A}$ *is nonsingular if* rank $\mathbf{A} = n$, *and is singular if* rank $\mathbf{A} < n$.

www.EngineeringEBooksPdf.com

PROOF Let $\mathbf{A}$ be a given $n \times n$ matrix and consider the linear system

$$\mathbf{Ax} = \mathbf{b}. \tag{2}$$

If the inverse $\mathbf{A}^{-1}$ exists, then multiplication from the left on both sides and use of (1) gives

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

This shows that (2) has a solution $\mathbf{x}$, which is unique because, for another solution $\mathbf{u}$, we have $\mathbf{Au} = \mathbf{b}$, so that $\mathbf{u} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{x}$. Hence $\mathbf{A}$ must have rank n by the Fundamental Theorem in Sec. 7.5.

Conversely, let rank $\mathbf{A} = n$. Then by the same theorem, the system (2) has a unique solution $\mathbf{x}$ for any $\mathbf{b}$. Now the back substitution following the Gauss elimination (Sec. 7.3) shows that the components x_j of $\mathbf{x}$ are linear combinations of those of $\mathbf{b}$. Hence we can write

$$\mathbf{x} = \mathbf{Bb} \tag{3}$$

with $\mathbf{B}$ to be determined. Substitution into (2) gives

$$\mathbf{Ax} = \mathbf{A}(\mathbf{Bb}) = (\mathbf{AB})\mathbf{b} = \mathbf{Cb} = \mathbf{b} \qquad (\mathbf{C} = \mathbf{AB})$$

for any $\mathbf{b}$. Hence $\mathbf{C} = \mathbf{AB} = \mathbf{I}$, the unit matrix. Similarly, if we substitute (2) into (3) we get

$$\mathbf{x} = \mathbf{Bb} = \mathbf{B}(\mathbf{Ax}) = (\mathbf{BA})\mathbf{x}$$

for any $\mathbf{x}$ (and $\mathbf{b} = \mathbf{Ax}$). Hence $\mathbf{BA} = \mathbf{I}$. Together, $\mathbf{B} = \mathbf{A}^{-1}$ exists. ■

Determination of the Inverse by the Gauss–Jordan Method

To actually determine the inverse $\mathbf{A}^{-1}$ of a nonsingular $n \times n$ matrix $\mathbf{A}$, we can use a variant of the Gauss elimination (Sec. 7.3), called the **Gauss–Jordan elimination**.[3] The idea of the method is as follows.

Using $\mathbf{A}$, we form n linear systems

$$\mathbf{Ax}_{(1)} = \mathbf{e}_{(1)}, \quad \cdots, \quad \mathbf{Ax}_{(n)} = \mathbf{e}_{(n)}$$

where the vectors $\mathbf{e}_{(1)}, \cdots, \mathbf{e}_{(n)}$ are the columns of the $n \times n$ unit matrix $\mathbf{I}$; thus, $\mathbf{e}_{(1)} = [1 \quad 0 \quad \cdots \quad 0]^\mathsf{T}$, $\mathbf{e}_{(2)} = [0 \quad 1 \quad 0 \quad \cdots \quad 0]^\mathsf{T}$, etc. These are n vector equations in the unknown vectors $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(n)}$. We combine them into a single matrix equation

[3] WILHELM JORDAN (1842–1899), German geodesist and mathematician. He did important geodesic work in Africa, where he surveyed oases. [See Althoen, S.C. and R. McLaughlin, Gauss–Jordan reduction: A brief history. *American Mathematical Monthly*, Vol. **94**, No. 2 (1987), pp. 130–142.]

We do ***not recommend*** it as a method for solving systems of linear equations, since the number of operations in addition to those of the Gauss elimination is larger than that for back substitution, which the Gauss–Jordan elimination avoids. See also Sec. 20.1.

www.EngineeringEBooksPdf.com

$\mathbf{AX} = \mathbf{I}$, with the unknown matrix $\mathbf{X}$ having the columns $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(n)}$. Correspondingly, we combine the n augmented matrices $[\mathbf{A} \quad \mathbf{e}_{(1)}], \cdots, [\mathbf{A} \quad \mathbf{e}_{(n)}]$ into one wide $n \times 2n$ "augmented matrix" $\tilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{I}]$. Now multiplication of $\mathbf{AX} = \mathbf{I}$ by $\mathbf{A}^{-1}$ from the left gives $\mathbf{X} = \mathbf{A}^{-1}\mathbf{I} = \mathbf{A}^{-1}$. Hence, to solve $\mathbf{AX} = \mathbf{I}$ for $\mathbf{X}$, we can apply the Gauss elimination to $\tilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{I}]$. This gives a matrix of the form $[\mathbf{U} \quad \mathbf{H}]$ with upper triangular $\mathbf{U}$ because the Gauss elimination triangularizes systems. The Gauss–Jordan method reduces $\mathbf{U}$ by further elementary row operations to diagonal form, in fact to the unit matrix $\mathbf{I}$. This is done by eliminating the entries of $\mathbf{U}$ above the main diagonal and making the diagonal entries all 1 by multiplication (see Example 1). Of course, the method operates on the entire matrix $[\mathbf{U} \quad \mathbf{H}]$, transforming $\mathbf{H}$ into some matrix $\mathbf{K}$, hence the entire $[\mathbf{U} \quad \mathbf{H}]$ to $[\mathbf{I} \quad \mathbf{K}]$. This is the "augmented matrix" of $\mathbf{IX} = \mathbf{K}$. Now $\mathbf{IX} = \mathbf{X} = \mathbf{A}^{-1}$, as shown before. By comparison, $\mathbf{K} = \mathbf{A}^{-1}$, so that we can read $\mathbf{A}^{-1}$ directly from $[\mathbf{I} \quad \mathbf{K}]$.

The following example illustrates the practical details of the method.

EXAMPLE 1 Finding the Inverse of a Matrix by Gauss–Jordan Elimination

Determine the inverse $\mathbf{A}^{-1}$ of

$$\mathbf{A} = \begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix}.$$

Solution. We apply the Gauss elimination (Sec. 7.3) to the following $n \times 2n = 3 \times 6$ matrix, where BLUE always refers to the previous matrix.

$$[\mathbf{A} \quad \mathbf{I}] = \left[\begin{array}{ccc|ccc} -1 & 1 & 2 & 1 & 0 & 0 \\ 3 & -1 & 1 & 0 & 1 & 0 \\ -1 & 3 & 4 & 0 & 0 & 1 \end{array}\right]$$

$$\left[\begin{array}{ccc|ccc} -1 & 1 & 2 & 1 & 0 & 0 \\ 0 & 2 & 7 & 3 & 1 & 0 \\ 0 & 2 & 2 & -1 & 0 & 1 \end{array}\right] \begin{array}{l} \\ \text{Row 2} + 3\,\text{Row 1} \\ \text{Row 3} - \text{Row 1} \end{array}$$

$$\left[\begin{array}{ccc|ccc} -1 & 1 & 2 & 1 & 0 & 0 \\ 0 & 2 & 7 & 3 & 1 & 0 \\ 0 & 0 & -5 & -4 & -1 & 1 \end{array}\right] \begin{array}{l} \\ \\ \text{Row 3} - \text{Row 2} \end{array}$$

This is $[\mathbf{U} \quad \mathbf{H}]$ as produced by the Gauss elimination. Now follow the additional Gauss–Jordan steps, reducing $\mathbf{U}$ to $\mathbf{I}$, that is, to diagonal form with entries 1 on the main diagonal.

$$\left[\begin{array}{ccc|ccc} 1 & -1 & -2 & -1 & 0 & 0 \\ 0 & 1 & 3.5 & 1.5 & 0.5 & 0 \\ 0 & 0 & 1 & 0.8 & 0.2 & -0.2 \end{array}\right] \begin{array}{l} -\text{Row 1} \\ 0.5\,\text{Row 2} \\ -0.2\,\text{Row 3} \end{array}$$

$$\left[\begin{array}{ccc|ccc} 1 & -1 & 0 & 0.6 & 0.4 & -0.4 \\ 0 & 1 & 0 & -1.3 & -0.2 & 0.7 \\ 0 & 0 & 1 & 0.8 & 0.2 & -0.2 \end{array}\right] \begin{array}{l} \text{Row 1} + 2\,\text{Row 3} \\ \text{Row 2} - 3.5\,\text{Row 3} \\ \\ \end{array}$$

$$\left[\begin{array}{ccc|ccc} 1 & 0 & 0 & -0.7 & 0.2 & 0.3 \\ 0 & 1 & 0 & -1.3 & -0.2 & 0.7 \\ 0 & 0 & 1 & 0.8 & 0.2 & -0.2 \end{array}\right] \begin{array}{l} \text{Row 1} + \text{Row 2} \\ \\ \\ \end{array}$$

www.EngineeringEBooksPdf.com

The last three columns constitute $\mathbf{A}^{-1}$. Check:

$$\begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix} \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Hence $\mathbf{AA}^{-1} = \mathbf{I}$. Similarly, $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. ■

Formulas for Inverses

Since finding the inverse of a matrix is really a problem of solving a system of linear equations, it is not surprising that Cramer's rule (Theorem 4, Sec. 7.7) might come into play. And similarly, as Cramer's rule was useful for theoretical study but not for computation, so too is the explicit formula (4) in the following theorem useful for theoretical considerations but not recommended for actually determining inverse matrices, except for the frequently occurring 2×2 case as given in (4*).

THEOREM 2

Inverse of a Matrix by Determinants

The inverse of a nonsingular $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is given by

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}}[C_{jk}]^{\mathsf{T}} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \cdot & \cdot & \cdots & \cdot \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{bmatrix}, \tag{4}$$

where C_{jk} is the cofactor of a_{jk} in det $\mathbf{A}$ (see Sec. 7.7). (CAUTION! Note well that in $\mathbf{A}^{-1}$, the cofactor C_{jk} occupies the same place as a_{kj} (not a_{jk}) does in $\mathbf{A}$.)

In particular, the inverse of

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \textit{is} \quad \mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \tag{4*}$$

PROOF We denote the right side of (4) by $\mathbf{B}$ and show that $\mathbf{BA} = \mathbf{I}$. We first write

$$\mathbf{BA} = \mathbf{G} = [g_{kl}] \tag{5}$$

and then show that $\mathbf{G} = \mathbf{I}$. Now by the definition of matrix multiplication and because of the form of $\mathbf{B}$ in (4), we obtain (CAUTION! C_{sk}, not C_{ks})

$$g_{kl} = \sum_{s=1}^{n} \frac{C_{sk}}{\det \mathbf{A}} a_{sl} = \frac{1}{\det \mathbf{A}} (a_{1l}C_{1k} + \cdots + a_{nl}C_{nk}). \tag{6}$$

www.EngineeringEBooksPdf.com

Now (9) and (10) in Sec. 7.7 show that the sum $(\cdots)$ on the right is $D = \det \mathbf{A}$ when $l = k$, and is zero when $l \neq k$. Hence

$$g_{kk} = \frac{1}{\det \mathbf{A}} \det \mathbf{A} = 1,$$

$$g_{kl} = 0 \quad (l \neq k).$$

In particular, for $n = 2$ we have in (4), in the first row, $C_{11} = a_{22}$, $C_{21} = -a_{12}$ and, in the second row, $C_{12} = -a_{21}$, $C_{22} = a_{11}$. This gives (4*). ■

The special case $n = 2$ occurs quite frequently in geometric and other applications. You may perhaps want to memorize formula (4*). Example 2 gives an illustration of (4*).

EXAMPLE 2 Inverse of a 2 × 2 Matrix by Determinants

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix}, \quad \mathbf{A}^{-1} = \frac{1}{10}\begin{bmatrix} 4 & -1 \\ -2 & 3 \end{bmatrix} = \begin{bmatrix} 0.4 & -0.1 \\ -0.2 & 0.3 \end{bmatrix}$$ ■

EXAMPLE 3 Further Illustration of Theorem 2

Using (4), find the inverse of

$$\mathbf{A} = \begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix}.$$

Solution. We obtain $\det \mathbf{A} = -1(-7) - 1 \cdot 13 + 2 \cdot 8 = 10$, and in (4),

$$C_{11} = \begin{vmatrix} -1 & 1 \\ 3 & 4 \end{vmatrix} = -7, \qquad C_{21} = -\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = 2, \qquad C_{31} = \begin{vmatrix} 1 & 2 \\ -1 & 1 \end{vmatrix} = 3,$$

$$C_{12} = -\begin{vmatrix} 3 & 1 \\ -1 & 4 \end{vmatrix} = -13, \quad C_{22} = \begin{vmatrix} -1 & 2 \\ -1 & 4 \end{vmatrix} = -2, \quad C_{32} = -\begin{vmatrix} -1 & 2 \\ 3 & 1 \end{vmatrix} = 7,$$

$$C_{13} = \begin{vmatrix} 3 & -1 \\ -1 & 3 \end{vmatrix} = 8, \qquad C_{23} = -\begin{vmatrix} -1 & 1 \\ -1 & 3 \end{vmatrix} = 2, \quad C_{33} = \begin{vmatrix} -1 & 1 \\ 3 & -1 \end{vmatrix} = -2,$$

so that by (4), in agreement with Example 1,

$$\mathbf{A}^{-1} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix}.$$ ■

Diagonal matrices $\mathbf{A} = [a_{jk}]$, $a_{jk} = 0$ when $j \neq k$, have an inverse if and only if all $a_{jj} \neq 0$. Then $\mathbf{A}^{-1}$ is diagonal, too, with entries $1/a_{11}, \cdots, 1/a_{nn}$.

PROOF For a diagonal matrix we have in (4)

$$\frac{C_{11}}{D} = \frac{a_{22} \cdots a_{nn}}{a_{11}a_{22} \cdots a_{nn}} = \frac{1}{a_{11}}, \qquad \text{etc.}$$ ■

www.EngineeringEBooksPdf.com

EXAMPLE 4 Inverse of a Diagonal Matrix

Let

$$\mathbf{A} = \begin{bmatrix} -0.5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then we obtain the inverse $\mathbf{A}^{-1}$ by inverting each individual diagonal element of $\mathbf{A}$, that is, by taking $1/(-0.5)$, $\frac{1}{4}$, and $\frac{1}{1}$ as the diagonal entries of $\mathbf{A}^{-1}$, that is,

$$\mathbf{A}^{-1} = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

■

Products can be inverted by taking the inverse of each factor and multiplying these inverses ***in reverse order***,

$$(\mathbf{AC})^{-1} = \mathbf{C}^{-1}\mathbf{A}^{-1}. \tag{7}$$

Hence for more than two factors,

$$(\mathbf{AC} \cdots \mathbf{PQ})^{-1} = \mathbf{Q}^{-1}\mathbf{P}^{-1} \cdots \mathbf{C}^{-1}\mathbf{A}^{-1}. \tag{8}$$

PROOF The idea is to start from (1) for $\mathbf{AC}$ instead of $\mathbf{A}$, that is, $\mathbf{AC}(\mathbf{AC})^{-1} = \mathbf{I}$, and multiply it on both sides from the left, first by $\mathbf{A}^{-1}$, which because of $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ gives

$$\begin{aligned} \mathbf{A}^{-1}\mathbf{AC}(\mathbf{AC})^{-1} &= \mathbf{C}(\mathbf{AC})^{-1} \\ &= \mathbf{A}^{-1}\mathbf{I} = \mathbf{A}^{-1}, \end{aligned}$$

and then multiplying this on both sides from the left, this time by $\mathbf{C}^{-1}$ and by using $\mathbf{C}^{-1}\mathbf{C} = \mathbf{I}$,

$$\mathbf{C}^{-1}\mathbf{C}(\mathbf{AC})^{-1} = (\mathbf{AC})^{-1} = \mathbf{C}^{-1}\mathbf{A}^{-1}.$$

This proves (7), and from it, (8) follows by induction. ■

We also note that *the inverse of the inverse is the given matrix*, as you may prove,

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}. \tag{9}$$

Unusual Properties of Matrix Multiplication. Cancellation Laws

Section 7.2 contains warnings that some properties of matrix multiplication deviate from those for numbers, and we are now able to explain the restricted validity of the so-called **cancellation laws** [2] and [3] below, using rank and inverse, concepts that were not yet

www.EngineeringEBooksPdf.com

available in Sec. 7.2. The deviations from the usual are of great practical importance and must be carefully observed. They are as follows.

[1] Matrix multiplication is not commutative, that is, in general we have

$$\mathbf{AB} \neq \mathbf{BA}.$$

[2] $\mathbf{AB} = \mathbf{0}$ does not generally imply $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$ (or $\mathbf{BA} = \mathbf{0}$); for example,

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

[3] $\mathbf{AC} = \mathbf{AD}$ does not generally imply $\mathbf{C} = \mathbf{D}$ (even when $\mathbf{A} \neq \mathbf{0}$).

Complete answers to [2] and [3] are contained in the following theorem.

THEOREM 3

Cancellation Laws

Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ *be* $n \times n$ *matrices. Then:*

(a) *If* rank $\mathbf{A} = n$ *and* $\mathbf{AB} = \mathbf{AC}$, *then* $\mathbf{B} = \mathbf{C}$.

(b) *If* rank $\mathbf{A} = n$, *then* $\mathbf{AB} = \mathbf{0}$ *implies* $\mathbf{B} = \mathbf{0}$. *Hence if* $\mathbf{AB} = \mathbf{0}$, *but* $\mathbf{A} \neq \mathbf{0}$ *as well as* $\mathbf{B} \neq \mathbf{0}$, *then* rank $\mathbf{A} < n$ *and* rank $\mathbf{B} < n$.

(c) *If* $\mathbf{A}$ *is singular, so are* $\mathbf{BA}$ *and* $\mathbf{AB}$.

PROOF **(a)** The inverse of $\mathbf{A}$ exists by Theorem 1. Multiplication by $\mathbf{A}^{-1}$ from the left gives $\mathbf{A}^{-1}\mathbf{AB} = \mathbf{A}^{-1}\mathbf{AC}$, hence $\mathbf{B} = \mathbf{C}$.

(b) Let rank $\mathbf{A} = n$. Then $\mathbf{A}^{-1}$ exists, and $\mathbf{AB} = \mathbf{0}$ implies $\mathbf{A}^{-1}\mathbf{AB} = \mathbf{B} = \mathbf{0}$. Similarly when rank $\mathbf{B} = n$. This implies the second statement in (b).

(c_1) Rank $\mathbf{A} < n$ by Theorem 1. Hence $\mathbf{Ax} = \mathbf{0}$ has nontrivial solutions by Theorem 2 in Sec. 7.5. Multiplication by $\mathbf{B}$ shows that these solutions are also solutions of $\mathbf{BAx} = \mathbf{0}$, so that rank $(\mathbf{BA}) < n$ by Theorem 2 in Sec. 7.5 and $\mathbf{BA}$ is singular by Theorem 1.

(c_2) $\mathbf{A}^\mathsf{T}$ is singular by Theorem 2(d) in Sec. 7.7. Hence $\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$ is singular by part (c_1), and is equal to $(\mathbf{AB})^\mathsf{T}$ by (10d) in Sec. 7.2. Hence $\mathbf{AB}$ is singular by Theorem 2(d) in Sec. 7.7. ■

Determinants of Matrix Products

The determinant of a matrix product $\mathbf{AB}$ or $\mathbf{BA}$ can be written as the product of the determinants of the factors, and it is interesting that det $\mathbf{AB}$ = det $\mathbf{BA}$, although $\mathbf{AB} \neq \mathbf{BA}$ in general. The corresponding formula (10) is needed occasionally and can be obtained by Gauss–Jordan elimination (see Example 1) and from the theorem just proved.

THEOREM 4

Determinant of a Product of Matrices

For any $n \times n$ *matrices* $\mathbf{A}$ *and* $\mathbf{B}$,

$$\det(\mathbf{AB}) = \det(\mathbf{BA}) = \det \mathbf{A} \det \mathbf{B}. \tag{10}$$

www.EngineeringEBooksPdf.com

PROOF If **A** or **B** is singular, so are **AB** and **BA** by Theorem 3(c), and (10) reduces to $0 = 0$ by Theorem 3 in Sec. 7.7.

Now let **A** and **B** be nonsingular. Then we can reduce **A** to a diagonal matrix $\hat{\mathbf{A}} = [a_{jk}]$ by Gauss–Jordan steps. Under these operations, det **A** retains its value, by Theorem 1 in Sec. 7.7, (a) and (b) [not (c)] except perhaps for a sign reversal in row interchanging when pivoting. But the same operations reduce **AB** to $\hat{\mathbf{A}}\mathbf{B}$ with the same effect on det (**AB**). Hence it remains to prove (10) for $\hat{\mathbf{A}}\mathbf{B}$; written out,

$$\hat{\mathbf{A}}\mathbf{B} = \begin{bmatrix} \hat{a}_{11} & 0 & \cdots & 0 \\ 0 & \hat{a}_{22} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \hat{a}_{nn} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ & & \vdots & \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} \hat{a}_{11}b_{11} & \hat{a}_{11}b_{12} & \cdots & \hat{a}_{11}b_{1n} \\ \hat{a}_{22}b_{21} & \hat{a}_{22}b_{22} & \cdots & \hat{a}_{22}b_{2n} \\ & & \vdots & \\ \hat{a}_{nn}b_{n1} & \hat{a}_{nn}b_{n2} & \cdots & \hat{a}_{nn}b_{nn} \end{bmatrix}.$$

We now take the determinant det ($\hat{\mathbf{A}}\mathbf{B}$). On the right we can take out a factor $\hat{a}_{11}$ from the first row, $\hat{a}_{22}$ from the second, $\cdots$, $\hat{a}_{nn}$ from the nth. But this product $\hat{a}_{11}\hat{a}_{22}\cdots\hat{a}_{nn}$ equals det $\hat{\mathbf{A}}$ because $\hat{\mathbf{A}}$ is diagonal. The remaining determinant is det **B**. This proves (10) for det (**AB**), and the proof for det (**BA**) follows by the same idea. ■

This completes our discussion of linear systems (Secs. 7.3–7.8). Section 7.9 on vector spaces and linear transformations is optional. ***Numeric methods*** are discussed in Secs. 20.1–20.4, which are independent of other sections on numerics.

PROBLEM SET 7.8

1–10 INVERSE

Find the inverse by Gauss–Jordan (or by (4*) if $n = 2$). Check by using (1).

1. $\begin{bmatrix} 1.80 & -2.32 \\ -0.25 & 0.60 \end{bmatrix}$

2. $\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ -\sin 2\theta & \cos 2\theta \end{bmatrix}$

3. $\begin{bmatrix} 0.3 & -0.1 & 0.5 \\ 2 & 6 & 4 \\ 5 & 0 & 9 \end{bmatrix}$

4. $\begin{bmatrix} 0 & 0 & 0.1 \\ 0 & -0.4 & 0 \\ 2.5 & 0 & 0 \end{bmatrix}$

5. $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 5 & 4 & 1 \end{bmatrix}$

6. $\begin{bmatrix} -4 & 0 & 0 \\ 0 & 8 & 13 \\ 0 & 3 & 5 \end{bmatrix}$

7. $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

8. $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$

9. $\begin{bmatrix} 0 & 8 & 0 \\ 0 & 0 & 4 \\ 2 & 0 & 0 \end{bmatrix}$

10. $\begin{bmatrix} \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \end{bmatrix}$

11–18 SOME GENERAL FORMULAS

11. Inverse of the square. Verify $(\mathbf{A}^2)^{-1} = (\mathbf{A}^{-1})^2$ for **A** in Prob. 1.

12. Prove the formula in Prob. 11.

www.EngineeringEBooksPdf.com

13. Inverse of the transpose. Verify $(\mathbf{A}^{\mathsf{T}})^{-1} = (\mathbf{A}^{-1})^{\mathsf{T}}$ for $\mathbf{A}$ in Prob. 1.

14. Prove the formula in Prob. 13.

15. Inverse of the inverse. Prove that $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.

16. Rotation. Give an application of the matrix in Prob. 2 that makes the form of the inverse obvious.

17. Triangular matrix. Is the inverse of a triangular matrix always triangular (as in Prob. 5)? Give reason.

18. Row interchange. Same task as in Prob. 16 for the matrix in Prob. 7.

19–20 FORMULA (4)

Formula (4) is occasionally needed in theory. To understand it, apply it and check the result by Gauss–Jordan:

19. In Prob. 3

20. In Prob. 6

7.9 Vector Spaces, Inner Product Spaces, Linear Transformations *Optional*

We have captured the essence of vector spaces in Sec. 7.4. There we dealt with ***special vector spaces*** that arose quite naturally in the context of matrices and linear systems. The elements of these vector spaces, called *vectors*, satisfied rules (3) and (4) of Sec. 7.1 (which were similar to those for numbers). These special vector spaces were generated by *spans,* that is, linear combination of finitely many vectors. Furthermore, each such vector had n real numbers as *components.* Review this material before going on.

We can generalize this idea by taking *all* vectors with n real numbers as components and obtain the very important ***real n-dimensional vector space*** R^n. The vectors are known as "real vectors." Thus, each vector in R^n is an ordered n-tuple of real numbers.

Now we can consider special values for n. For $n = 2$, we obtain R^2, the vector space of all ordered pairs, which correspond to the **vectors in the plane**. For $n = 3$, we obtain R^3, the vector space of all ordered triples, which are the **vectors in 3-space**. These vectors have wide applications in mechanics, geometry, and calculus and are basic to the engineer and physicist.

Similarly, if we take all ordered n-tuples of *complex numbers* as vectors and complex numbers as scalars, we obtain the **complex vector space** C^n, which we shall consider in Sec. 8.5.

Furthermore, there are other sets of practical interest consisting of matrices, functions, transformations, or others for which addition and scalar multiplication can be defined in an almost natural way so that they too form vector spaces.

It is perhaps not too great an intellectual jump to create, from the ***concrete model*** R^n, the ***abstract concept*** of a ***real vector space*** V by taking the basic properties (3) and (4) in Sec. 7.1 as axioms. In this way, the definition of a real vector space arises.

DEFINITION

Real Vector Space

A nonempty set V of elements $\mathbf{a}, \mathbf{b}, \cdots$ is called a **real vector space** (or *real linear space*), and these elements are called **vectors** (regardless of their nature, which will come out from the context or will be left arbitrary) if, in V, there are defined two algebraic operations (called *vector addition* and *scalar multiplication*) as follows.

I. Vector addition associates with every pair of vectors $\mathbf{a}$ and $\mathbf{b}$ of V a unique vector of V, called the *sum* of $\mathbf{a}$ and $\mathbf{b}$ and denoted by $\mathbf{a} + \mathbf{b}$, such that the following axioms are satisfied.

www.EngineeringEBooksPdf.com

I.1 *Commutativity.* For any two vectors **a** and **b** of V,

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}.$$

I.2 *Associativity.* For any three vectors **a**, **b**, **c** of V,

$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c}) \qquad \text{(written } \mathbf{a} + \mathbf{b} + \mathbf{c}\text{)}.$$

I.3 There is a unique vector in V, called the *zero vector* and denoted by **0**, such that for every **a** in V,

$$\mathbf{a} + \mathbf{0} = \mathbf{a}.$$

I.4 For every **a** in V there is a unique vector in V that is denoted by $-\mathbf{a}$ and is such that

$$\mathbf{a} + (-\mathbf{a}) = \mathbf{0}.$$

II. Scalar multiplication. The real numbers are called **scalars**. Scalar multiplication associates with every **a** in V and every scalar c a unique vector of V, called the *product* of c and **a** and denoted by $c\mathbf{a}$ (or $\mathbf{a}c$) such that the following axioms are satisfied.

II.1 *Distributivity.* For every scalar c and vectors **a** and **b** in V,

$$c(\mathbf{a} + \mathbf{b}) = c\mathbf{a} + c\mathbf{b}.$$

II.2 *Distributivity.* For all scalars c and k and every **a** in V,

$$(c + k)\mathbf{a} = c\mathbf{a} + k\mathbf{a}.$$

II.3 *Associativity.* For all scalars c and k and every **a** in V,

$$c(k\mathbf{a}) = (ck)\mathbf{a} \qquad \text{(written } ck\mathbf{a}\text{)}.$$

II.4 For every **a** in V,

$$1\mathbf{a} = \mathbf{a}.$$

If, in the above definition, we take complex numbers as scalars instead of real numbers, we obtain the axiomatic definition of a **complex vector space**.

Take a look at the axioms in the above definition. Each axiom stands on its own: It is concise, useful, and it expresses a simple property of V. There are as few axioms as possible and together they express ***all*** the desired properties of V. Selecting good axioms is a process of trial and error that often extends over a long period of time. But once agreed upon, axioms become *standard* such as the ones in the definition of a real vector space.

www.EngineeringEBooksPdf.com

The following concepts related to a vector space are exactly defined as those given in Sec. 7.4. Indeed, a **linear combination** of vectors $a_{(1)}, \cdots, a_{(m)}$ in a vector space V is an expression

$$c_1\mathbf{a}_{(1)} + \cdots + c_m\mathbf{a}_m \qquad (c_1, \cdots, c_m \text{ any scalars}).$$

These vectors form a **linearly independent set** (briefly, they are called **linearly independent**) if

$$c_1\mathbf{a}_{(1)} + \cdots + c_m\mathbf{a}_{(m)} = \mathbf{0} \tag{1}$$

implies that $c_1 = 0, \cdots, c_m = 0$. Otherwise, if (1) also holds with scalars not all zero, the vectors are called **linearly dependent**.

Note that (1) with $m = 1$ is $c\mathbf{a} = \mathbf{0}$ and shows that a single vector $\mathbf{a}$ is linearly independent if and only if $\mathbf{a} \neq \mathbf{0}$.

V has **dimension n**, or is **n-dimensional**, if it contains a linearly independent set of n vectors, whereas any set of more than n vectors in V is linearly dependent. That set of n linearly independent vectors is called a **basis** for V. Then every vector in V can be written as a linear combination of the basis vectors. Furthermore, for a given basis, this representation is unique (see Prob. 2).

EXAMPLE 1 Vector Space of Matrices

The real 2×2 matrices form a four-dimensional real vector space. A basis is

$$\mathbf{B}_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B}_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B}_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{B}_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

because any 2×2 matrix $\mathbf{A} = [a_{jk}]$ has a unique representation $\mathbf{A} = a_{11}\mathbf{B}_{11} + a_{12}\mathbf{B}_{12} + a_{21}\mathbf{B}_{21} + a_{22}\mathbf{B}_{22}$. Similarly, the real $m \times n$ matrices with fixed m and n form an mn-dimensional vector space. What is the dimension of the vector space of all 3×3 skew-symmetric matrices? Can you find a basis? ■

EXAMPLE 2 Vector Space of Polynomials

The set of all constant, linear, and quadratic polynomials in x together is a vector space of dimension 3 with basis $\{1, x, x^2\}$ under the usual addition and multiplication by real numbers because these two operations give polynomials not exceeding degree 2. What is the dimension of the vector space of all polynomials of degree not exceeding a given fixed n? Can you find a basis? ■

If a vector space V contains a linearly independent set of n vectors for every n, no matter how large, then V is called **infinite dimensional**, as opposed to a *finite dimensional* (n-dimensional) vector space just defined. An example of an infinite dimensional vector space is the space of all continuous functions on some interval $[a, b]$ of the x-axis, as we mention without proof.

Inner Product Spaces

If $\mathbf{a}$ and $\mathbf{b}$ are vectors in R^n, regarded as column vectors, we can form the product $\mathbf{a}^\mathsf{T}\mathbf{b}$. This is a 1×1 matrix, which we can identify with its single entry, that is, with a number.

www.EngineeringEBooksPdf.com

This product is called the **inner product** or **dot product** of **a** and **b**. Other notations for it are (**a**, **b**) and **a** • **b**. Thus

$$\mathbf{a}^{\mathsf{T}}\mathbf{b} = (\mathbf{a}, \mathbf{b}) = \mathbf{a} \bullet \mathbf{b} = [a_1 \cdots a_n] \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \sum_{i=1}^{n} a_l b_l = a_1 b_1 + \cdots + a_n b_n.$$

We now extend this concept to general real vector spaces by taking basic properties of (**a**, **b**) as axioms for an "abstract inner product" (**a**, **b**) as follows.

DEFINITION

Real Inner Product Space

A real vector space V is called a **real inner product space** (or *real pre-Hilbert*[4] *space*) if it has the following property. With every pair of vectors **a** and **b** in V there is associated a real number, which is denoted by (**a**, **b**) and is called the **inner product** of **a** and **b**, such that the following axioms are satisfied.

I. For all scalars q_1 and q_2 and all vectors **a**, **b**, **c** in V,

$$(q_1\mathbf{a} + q_2\mathbf{b}, \mathbf{c}) = q_1(\mathbf{a}, \mathbf{c}) + q_2(\mathbf{b}, \mathbf{c}) \qquad (\textit{Linearity}).$$

II. For all vectors **a** and **b** in V,

$$(\mathbf{a}, \mathbf{b}) = (\mathbf{b}, \mathbf{a}) \qquad (\textit{Symmetry}).$$

III. For every **a** in V,

$$\left.\begin{array}{r} (\mathbf{a}, \mathbf{a}) \geqq 0, \\ (\mathbf{a}, \mathbf{a}) = 0 \quad \text{if and only if} \quad \mathbf{a} = \mathbf{0} \end{array}\right\} \quad (\textit{Positive-definiteness}).$$

Vectors whose inner product is zero are called **orthogonal**.
The *length* or **norm** of a vector in V is defined by

$$\|\mathbf{a}\| = \sqrt{(\mathbf{a}, \mathbf{a})} \quad (\geqq 0). \tag{2}$$

A vector of norm 1 is called a **unit vector**.

[4]DAVID HILBERT (1862–1943), great German mathematician, taught at Königsberg and Göttingen and was the creator of the famous Göttingen mathematical school. He is known for his basic work in algebra, the calculus of variations, integral equations, functional analysis, and mathematical logic. His "Foundations of Geometry" helped the axiomatic method to gain general recognition. His famous 23 problems (presented in 1900 at the International Congress of Mathematicians in Paris) considerably influenced the development of modern mathematics.

If V is finite dimensional, it is actually a so-called *Hilbert space*; see [GenRef7], p. 128, listed in App. 1.

www.EngineeringEBooksPdf.com

From these axioms and from (2) one can derive the basic inequality

$$|(\mathbf{a}, \mathbf{b})| \leqq \|\mathbf{a}\| \, \|\mathbf{b}\| \qquad (\textit{Cauchy–Schwarz}^5 \textit{ inequality}). \tag{3}$$

From this follows

$$\|\mathbf{a} + \mathbf{b}\| \leqq \|\mathbf{a}\| + \|\mathbf{b}\| \qquad (\textit{Triangle inequality}). \tag{4}$$

A simple direct calculation gives

$$\|\mathbf{a} + \mathbf{b}\|^2 + \|\mathbf{a} - \mathbf{b}\|^2 = 2(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2) \qquad (\textit{Parallelogram equality}). \tag{5}$$

EXAMPLE 3 ***n*-Dimensional Euclidean Space**

R^n with the inner product

$$(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\mathsf{T}\mathbf{b} = a_1 b_1 + \cdots + a_n b_n \tag{6}$$

(where both **a** and **b** are *column* vectors) is called the ***n*-dimensional Euclidean space** and is denoted by E^n or again simply by R^n. Axioms I–III hold, as direct calculation shows. Equation (2) gives the "**Euclidean norm**"

$$\|\mathbf{a}\| = \sqrt{(\mathbf{a}, \mathbf{a})} = \sqrt{\mathbf{a}^\mathsf{T}\mathbf{a}} = \sqrt{a_1^2 + \cdots + a_n^2}. \tag{7}$$

■

EXAMPLE 4 **An Inner Product for Functions. Function Space**

The set of all real-valued continuous functions $f(x)$, $g(x)$, $\cdots$ on a given interval $\alpha \leqq x \leqq \beta$ is a real vector space under the usual addition of functions and multiplication by scalars (real numbers). On this "**function space**" we can define an inner product by the integral

$$(f, g) = \int_\alpha^\beta f(x)\, g(x)\, dx. \tag{8}$$

Axioms I–III can be verified by direct calculation. Equation (2) gives the norm

$$\|f\| = \sqrt{(f, f)} = \sqrt{\int_\alpha^\beta f(x)^2\, dx}. \tag{9}$$

■

Our examples give a first impression of the great generality of the abstract concepts of vector spaces and inner product spaces. Further details belong to more advanced courses (on functional analysis, meaning abstract modern analysis; see [GenRef7] listed in App. 1) and cannot be discussed here. Instead we now take up a related topic where matrices play a central role.

Linear Transformations

Let X and Y be any vector spaces. To each vector **x** in X we assign a unique vector **y** in Y. Then we say that a **mapping** (or **transformation** or **operator**) of X into Y is given. Such a mapping is denoted by a capital letter, say F. The vector **y** in Y assigned to a vector **x** in X is called the **image** of **x** under F and is denoted by $F(x)$ [or $F\mathbf{x}$, without parentheses].

[5]HERMANN AMANDUS SCHWARZ (1843–1921). German mathematician, known by his work in complex analysis (conformal mapping) and differential geometry. For Cauchy see Sec. 2.5.

www.EngineeringEBooksPdf.com

F is called a **linear mapping** or **linear transformation** if, for all vectors $\mathbf{v}$ and $\mathbf{x}$ in X and scalars c,

$$F(\mathbf{v} + \mathbf{x}) = F(\mathbf{v}) + F(\mathbf{x}) \qquad F(c\mathbf{x}) = cF(\mathbf{x}). \tag{10}$$

Linear Transformation of Space R^n into Space R^m

From now on we let $X = R^n$ and $Y = R^m$. Then any real $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ gives a transformation of R^n into R^m,

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{11}$$

Since $\mathbf{A}(\mathbf{u} + \mathbf{x}) = \mathbf{A}\mathbf{u} + \mathbf{A}\mathbf{x}$ and $\mathbf{A}(c\mathbf{x}) = c\mathbf{A}\mathbf{x}$, this transformation is linear.

We show that, conversely, every linear transformation F of R^n into R^m can be given in terms of an $m \times n$ matrix $\mathbf{A}$, after a basis for R^n and a basis for R^m have been chosen. This can be proved as follows.

Let $\mathbf{e}_{(1)}, \cdots, \mathbf{e}_{(n)}$ be any basis for R^n. Then every $\mathbf{x}$ in R^n has a unique representation

$$\mathbf{x} = x_1\mathbf{e}_{(1)} + \cdots + x_n\mathbf{e}_{(n)}.$$

Since F is linear, this representation implies for the image $F(\mathbf{x})$:

$$F(\mathbf{x}) = F(x_1\mathbf{e}_{(1)} + \cdots + x_n\mathbf{e}_{(n)}) = x_1F(\mathbf{e}_{(1)}) + \cdots + x_nF(\mathbf{e}_{(n)}).$$

Hence F is uniquely determined by the images of the vectors of a basis for R^n. We now choose for R^n the "**standard basis**"

$$\mathbf{e}_{(1)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{e}_{(2)} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \cdots, \quad \mathbf{e}_{(n)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \tag{12}$$

where $\mathbf{e}_{(j)}$ has its jth component equal to 1 and all others 0. We show that we can now determine an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ such that for every $\mathbf{x}$ in R^n and image $\mathbf{y} = F(\mathbf{x})$ in R^m,

$$\mathbf{y} = F(\mathbf{x}) = \mathbf{A}\mathbf{x}.$$

Indeed, from the image $\mathbf{y}^{(1)} = F(\mathbf{e}_{(1)})$ of $\mathbf{e}_{(1)}$ we get the condition

$$\mathbf{y}^{(1)} = \begin{bmatrix} y_1^{(1)} \\ y_2^{(1)} \\ \vdots \\ y_m^{(1)} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

www.EngineeringEBooksPdf.com

from which we can determine the first column of $\mathbf{A}$, namely $a_{11} = y_1^{(1)}$, $a_{21} = y_2^{(1)}$, $\cdots$, $a_{m1} = y_m^{(1)}$. Similarly, from the image of $\mathbf{e}_{(2)}$ we get the second column of $\mathbf{A}$, and so on. This completes the proof. ■

We say that **A represents** F, or *is a representation of* F, with respect to the bases for R^n and R^m. Quite generally, the purpose of a "**representation**" is the replacement of one object of study by another object whose properties are more readily apparent.

In three-dimensional Euclidean space E^3 the standard basis is usually written $\mathbf{e}_{(1)} = \mathbf{i}$, $\mathbf{e}_{(2)} = \mathbf{j}$, $\mathbf{e}_{(3)} = \mathbf{k}$. Thus,

$$\mathbf{i} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{j} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{k} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{13}$$

These are the three unit vectors in the positive directions of the axes of the **Cartesian coordinate system in space**, that is, the usual coordinate system with the same scale of measurement on the three mutually perpendicular coordinate axes.

EXAMPLE 5 Linear Transformations

Interpreted as transformations of Cartesian coordinates in the plane, the matrices

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$$

represent a reflection in the line $x_2 = x_1$, a reflection in the x_1-axis, a reflection in the origin, and a stretch (when $a > 1$, or a contraction when $0 < a < 1$) in the x_1-direction, respectively. ■

EXAMPLE 6 Linear Transformations

Our discussion preceding Example 5 is simpler than it may look at first sight. To see this, find $\mathbf{A}$ representing the linear transformation that maps (x_1, x_2) onto $(2x_1 - 5x_2, 3x_1 + 4x_2)$.

Solution. Obviously, the transformation is

$$\begin{aligned} y_1 &= 2x_1 - 5x_2 \\ y_2 &= 3x_1 + 4x_2. \end{aligned}$$

From this we can directly see that the matrix is

$$\mathbf{A} = \begin{bmatrix} 2 & -5 \\ 3 & 4 \end{bmatrix}. \qquad \text{Check:} \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 & -5 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 - 5x_2 \\ 3x_1 + 4x_2 \end{bmatrix}.$$ ■

If $\mathbf{A}$ in (11) is square, $n \times n$, then (11) maps R^n into R^n. If this $\mathbf{A}$ is nonsingular, so that $\mathbf{A}^{-1}$ exists (see Sec. 7.8), then multiplication of (11) by $\mathbf{A}^{-1}$ from the left and use of $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ gives the **inverse transformation**

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}. \tag{14}$$

It maps every $\mathbf{y} = \mathbf{y}_0$ onto that $\mathbf{x}$, which by (11) is mapped onto $\mathbf{y}_0$. *The inverse of a linear transformation is itself linear,* because it is given by a matrix, as (14) shows.

www.EngineeringEBooksPdf.com

Composition of Linear Transformations

We want to give you a flavor of how linear transformations in general vector spaces work. You will notice, if you read carefully, that definitions and verifications (Example 7) strictly follow the given rules and you can think your way through the material by going in a slow systematic fashion.

The last operation we want to discuss is composition of linear transformations. Let X, Y, W be general vector spaces. As before, let F be a linear transformation from X to Y. Let G be a linear transformation from W to X. Then we denote, by H, the **composition** of F and G, that is,

$$H = F \circ G = FG = F(G),$$

which means *we take transformation G and then apply transformation F to it* (***in that order!***, *i.e. you go from left to right*).

Now, to give this a more concrete meaning, if we let $\mathbf{w}$ be a vector in W, then $G(\mathbf{w})$ is a vector in X and $F(G(\mathbf{w}))$ is a vector in Y. Thus, H maps W to Y, and we can write

$$H(\mathbf{w}) = (F \circ G)(\mathbf{w}) = (FG)(\mathbf{w}) = F(G(\mathbf{w})), \tag{15}$$

which completes the definition of composition in a general vector space setting. But is composition really linear? To check this we have to verify that H, as defined in (15), obeys the two equations of (10).

EXAMPLE 7 The Composition of Linear Transformations Is Linear

To show that H is indeed linear we must show that (10) holds. We have, for two vectors $\mathbf{w}_1$, $\mathbf{w}_2$ in W,

$$\begin{aligned}
H(\mathbf{w}_1 + \mathbf{w}_2) &= (F \circ G)(\mathbf{w}_1 + \mathbf{w}_2) \\
&= F(G(\mathbf{w}_1 + \mathbf{w}_2)) \\
&= F(G(\mathbf{w}_1) + G(\mathbf{w}_2)) && \text{(by linearity of } G\text{)} \\
&= F(G(\mathbf{w}_1)) + F(G(\mathbf{w}_2)) && \text{(by linearity of } F\text{)} \\
&= (F \circ G)(\mathbf{w}_1) + (F \circ G)(\mathbf{w}_2) && \text{(by (15))} \\
&= H(\mathbf{w}_1) + H(\mathbf{w}_2) && \text{(by definition of } H\text{).}
\end{aligned}$$

Similarly, $H(c\mathbf{w}_2) = (F \circ G)(c\mathbf{w}_2) = F(G(c\mathbf{w}_2)) = F(c(G(\mathbf{w}_2))$

$$= cF(G(\mathbf{w}_2)) = c(F \circ G)(\mathbf{w}_2) = cH(\mathbf{w}_2).$$

■

We defined composition as a linear transformation in a general vector space setting and showed that the composition of linear transformations is indeed linear.

Next we want to relate composition of linear transformations to matrix multiplication.

To do so we let $X = R^n$, $Y = R^m$, and $W = R^p$. This choice of particular vector spaces allows us to represent the linear transformations as matrices and form matrix equations, as was done in (11). Thus F can be represented by a general real $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ and G by an $n \times p$ matrix $\mathbf{B} = [b_{jk}]$. Then we can write for F, with column vectors $\mathbf{x}$ with n entries, and resulting vector $\mathbf{y}$, with m entries

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{16}$$

www.EngineeringEBooksPdf.com

and similarly for G, with column vector $\mathbf{w}$ with p entries,

$$\mathbf{x} = \mathbf{B}\mathbf{w}. \tag{17}$$

Substituting (17) into (16) gives

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{B}\mathbf{w}) = (\mathbf{A}\mathbf{B})\mathbf{w} = \mathbf{A}\mathbf{B}\mathbf{w} = \mathbf{C}\mathbf{w} \qquad \text{where } \mathbf{C} = \mathbf{A}\mathbf{B}. \tag{18}$$

This is (15) in a matrix setting, this is, ***we can define the composition of linear transformations in the Euclidean spaces as multiplication by matrices***. Hence, the real $m \times p$ matrix $\mathbf{C}$ represents a linear transformation H which maps R^p to R^n with vector $\mathbf{w}$, a column vector with p entries.

Remarks. Our discussion is similar to the one in Sec. 7.2, where we motivated the "unnatural" matrix multiplication of matrices. Look back and see that our current, more general, discussion is written out there for the case of dimension $m = 2$, $n = 2$, and $p = 2$. (You may want to write out our development by picking small *distinct* dimensions, such as $m = 2$, $n = 3$, and $p = 4$, and writing down the matrices and vectors. This is a trick of the trade of mathematicians in that we like to develop and test theories on smaller examples to see that they work.)

EXAMPLE 8 Linear Transformations. Composition

In Example 5 of Sec. 7.9, let $\mathbf{A}$ be the first matrix and $\mathbf{B}$ be the fourth matrix with $a > 1$. Then, applying $\mathbf{B}$ to a vector $\mathbf{w} = [w_1 \quad w_2]^\mathsf{T}$, stretches the element w_1 by a in the x_1 direction. Next, when we apply $\mathbf{A}$ to the "stretched" vector, we reflect the vector along the line $x_1 = x_2$, resulting in a vector $\mathbf{y} = [w_2 \quad aw_1]^\mathsf{T}$. But this represents, precisely, a geometric description for the composition H of two linear transformations F and G represented by matrices $\mathbf{A}$ and $\mathbf{B}$. We now show that, for this example, our result can be obtained by straightforward matrix multiplication, that is,

$$\mathbf{AB} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix}$$

and as in (18) calculate

$$\mathbf{ABw} = \begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} w_2 \\ aw_1 \end{bmatrix},$$

which is the same as before. This shows that indeed $\mathbf{AB} = \mathbf{C}$, and we see the composition of linear transformations can be represented by a linear transformation. It also shows that the order of matrix multiplication is important (!). You may want to try applying $\mathbf{A}$ first and then $\mathbf{B}$, resulting in $\mathbf{BA}$. What do you see? Does it make geometric sense? Is it the same result as $\mathbf{AB}$? ■

We have learned several abstract concepts such as vector space, inner product space, and linear transformation. *The introduction of such concepts allows engineers and scientists to communicate in a concise and common language.* For example, the concept of a vector space encapsulated a lot of ideas in a very concise manner. For the student, learning such concepts provides a foundation for more advanced studies in engineering.

This concludes Chapter 7. The central theme was *the Gaussian elimination of* Sec. 7.3 from which most of the other concepts and theory flowed. The next chapter again has a central theme, that is, *eigenvalue problems,* an area very rich in applications such as in engineering, modern physics, and other areas.

www.EngineeringEBooksPdf.com

PROBLEM SET 7.9

1. **Basis.** Find three bases of R^2.
2. **Uniqueness.** Show that the representation $\mathbf{v} = c_1\mathbf{a}_{(1)} + \cdots + c_n\mathbf{a}_{(n)}$ of any given vector in an n-dimensional vector space V in terms of a given basis $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(n)}$ for V is unique. ***Hint***. Take two representations and consider the difference.

3–10 VECTOR SPACE

(More problems in Problem Set 9.4.) Is the given set, taken with the usual addition and scalar multiplication, a vector space? Give reason. If your answer is yes, find the dimension and a basis.

3. All vectors in R^3 satisfying $-v_1 + 2v_2 + 3v_3 = 0$, $-4v_1 + v_2 + v_3 = 0$.
4. All skew-symmetric 3×3 matrices.
5. All polynomials in x of degree 4 or less with nonnegative coefficients.
6. All functions $y(x) = a \cos 2x + b \sin 2x$ with arbitrary constants a and b.
7. All functions $y(x) = (ax + b)e^{-x}$ with any constant a and b.
8. All $n \times n$ matrices $\mathbf{A}$ with fixed n and $\det \mathbf{A} = 0$.
9. All 2×2 matrices $[a_{jk}]$ with $a_{11} + a_{22} = 0$.
10. All 3×2 matrices $[a_{jk}]$ with first column any multiple of $[3 \quad 0 \quad -5]^\mathsf{T}$.

11–14 LINEAR TRANSFORMATIONS

Find the inverse transformation. Show the details.

11. $y_1 = 0.5x_1 - 0.5x_2$
    $y_2 = 1.5x_1 - 2.5x_2$

12. $y_1 = 3x_1 + 2x_2$
    $y_2 = 4x_1 + x_2$

13. $y_1 = 5x_1 + 3x_2 - 3x_3$
    $y_2 = 3x_1 + 2x_2 - 2x_3$
    $y_3 = 2x_1 - x_2 + 2x_3$

14. $y_1 = 0.2x_1 - 0.1x_2$
    $y_2 = -0.2x_2 + 0.1x_3$
    $y_3 = 0.1x_1 + 0.1x_3$

15–20 EUCLIDEAN NORM

Find the Euclidean norm of the vectors:

15. $[3 \quad 1 \quad -4]^\mathsf{T}$
16. $[\frac{1}{2} \quad \frac{1}{3} \quad -\frac{1}{2} \quad -\frac{1}{3}]^\mathsf{T}$
17. $[1 \quad 0 \quad 0 \quad 1 \quad -1 \quad 0 \quad -1 \quad 1]^\mathsf{T}$
18. $[-4 \quad 8 \quad -1]^\mathsf{T}$
19. $[\frac{2}{3} \quad \frac{2}{3} \quad \frac{1}{3} \quad 0]^\mathsf{T}$
20. $[\frac{1}{2} \quad -\frac{1}{2} \quad -\frac{1}{2} \quad \frac{1}{2}]^\mathsf{T}$

21–25 INNER PRODUCT. ORTHOGONALITY

21. **Orthogonality.** For what value(s) of k are the vectors $[2 \quad \frac{1}{2} \quad -4 \quad 0]^\mathsf{T}$ and $[5 \quad k \quad 0 \quad \frac{1}{4}]^\mathsf{T}$ orthogonal?
22. **Orthogonality.** Find all vectors in R^3 orthogonal to $[2 \quad 0 \quad 1]$. Do they form a vector space?
23. **Triangle inequality.** Verify (4) for the vectors in Probs. 15 and 18.
24. **Cauchy–Schwarz inequality.** Verify (3) for the vectors in Probs. 16 and 19.
25. **Parallelogram equality.** Verify (5) for the first two column vectors of the coefficient matrix in Prob. 13.

CHAPTER 7 REVIEW QUESTIONS AND PROBLEMS

1. What properties of matrix multiplication differ from those of the multiplication of numbers?
2. Let $\mathbf{A}$ be a 100×100 matrix and $\mathbf{B}$ a 100×50 matrix. Are the following expressions defined or not? $\mathbf{A} + \mathbf{B}$, $\mathbf{A}^2$, $\mathbf{B}^2$, $\mathbf{AB}$, $\mathbf{BA}$, $\mathbf{AA}^\mathsf{T}$, $\mathbf{B}^\mathsf{T}\mathbf{A}$, $\mathbf{B}^\mathsf{T}\mathbf{B}$, $\mathbf{BB}^\mathsf{T}$, $\mathbf{B}^\mathsf{T}\mathbf{AB}$. Give reasons.
3. Are there any linear systems without solutions? With one solution? With more than one solution? Give simple examples.
4. Let $\mathbf{C}$ be 10×10 matrix and $\mathbf{a}$ a column vector with 10 components. Are the following expressions defined or not? $\mathbf{Ca}$, $\mathbf{C}^\mathsf{T}\mathbf{a}$, $\mathbf{Ca}^\mathsf{T}$, $\mathbf{aC}$, $\mathbf{a}^\mathsf{T}\mathbf{C}$, $(\mathbf{Ca}^\mathsf{T})^\mathsf{T}$.
5. Motivate the definition of matrix multiplication.
6. Explain the use of matrices in linear transformations.
7. How can you give the rank of a matrix in terms of row vectors? Of column vectors? Of determinants?
8. What is the role of rank in connection with solving linear systems?
9. What is the idea of Gauss elimination and back substitution?
10. What is the inverse of a matrix? When does it exist? How would you determine it?

www.EngineeringEBooksPdf.com

11–20 MATRIX AND VECTOR CALCULATIONS

Showing the details, calculate the following expressions or give reason why they are not defined, when

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & -3 \\ 1 & 4 & 2 \\ -3 & 2 & 5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 4 & 1 \\ -4 & 0 & -2 \\ -1 & 2 & 0 \end{bmatrix},$$

$$\mathbf{u} = \begin{bmatrix} 2 \\ 0 \\ -5 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} 7 \\ -3 \\ 3 \end{bmatrix}$$

11. $\mathbf{AB}$, $\mathbf{BA}$

12. $\mathbf{A}^\mathsf{T}$, $\mathbf{B}^\mathsf{T}$

13. $\mathbf{Au}$, $\mathbf{u}^\mathsf{T}\mathbf{A}$

14. $\mathbf{u}^\mathsf{T}\mathbf{v}$, $\mathbf{uv}^\mathsf{T}$

15. $\mathbf{u}^\mathsf{T}\mathbf{Au}$, $\mathbf{v}^\mathsf{T}\mathbf{Bv}$

16. $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$

17. $\det \mathbf{A}$, $\det \mathbf{A}^2$, $(\det \mathbf{A})^2$, $\det \mathbf{B}$

18. $(\mathbf{A}^2)^{-1}$, $(\mathbf{A}^{-1})^2$

19. $\mathbf{AB} - \mathbf{BA}$

20. $(\mathbf{A} + \mathbf{A}^\mathsf{T})(\mathbf{B} - \mathbf{B}^\mathsf{T})$

21–28 LINEAR SYSTEMS

Showing the details, find all solutions or indicate that no solution exists.

21.
$$\begin{aligned} 4y + z &= 0 \\ 12x - 5y - 3z &= 34 \\ -6x \qquad + 4z &= 8 \end{aligned}$$

22.
$$\begin{aligned} 5x - 3y + z &= 7 \\ 2x + 3y - z &= 0 \\ 8x + 9y - 3z &= 2 \end{aligned}$$

23.
$$\begin{aligned} 9x + 3y - 6z &= 60 \\ 2x - 4y + 8z &= 4 \end{aligned}$$

24.
$$\begin{aligned} -6x + 39y - 9z &= -12 \\ 2x - 13y + 3z &= 4 \end{aligned}$$

25.
$$\begin{aligned} 0.3x - 0.7y + 1.3z &= 3.24 \\ 0.9y - 0.8z &= -2.53 \\ 0.7z &= 1.19 \end{aligned}$$

26.
$$\begin{aligned} 2x + 3y - 7z &= 3 \\ -4x - 6y + 14z &= 7 \end{aligned}$$

27.
$$\begin{aligned} x + 2y &= 6 \\ 3x + 5y &= 20 \\ -4x + y &= -42 \end{aligned}$$

28.
$$\begin{aligned} -8x \qquad + 2z &= 1 \\ 6y + 4z &= 3 \\ 12x + 2y \qquad &= 2 \end{aligned}$$

29–32 RANK

Determine the ranks of the coefficient matrix and the augmented matrix and state how many solutions the linear system will have.

29. In Prob. 23

30. In Prob. 24

31. In Prob. 27

32. In Prob. 26

33–35 NETWORKS

Find the currents.

33.

34.

35.

www.EngineeringEBooksPdf.com

SUMMARY OF CHAPTER 7

Linear Algebra: Matrices, Vectors, Determinants. Linear Systems

An $m \times n$ **matrix** $\mathbf{A} = [a_{jk}]$ is a rectangular array of numbers or functions ("entries," "elements") arranged in m horizontal **rows** and n vertical **columns**. If $m = n$, the matrix is called **square**. A $1 \times n$ matrix is called a **row vector** and an $m \times 1$ matrix a **column vector** (Sec. 7.1).

The **sum** $\mathbf{A} + \mathbf{B}$ of matrices of the same **size** (i.e., both $m \times n$) is obtained by adding corresponding entries. The **product** of $\mathbf{A}$ by a scalar c is obtained by multiplying each a_{jk} by c (Sec. 7.1).

The **product** $\mathbf{C} = \mathbf{AB}$ of an $m \times n$ matrix $\mathbf{A}$ by an $r \times p$ matrix $\mathbf{B} = [b_{jk}]$ is defined only when $r = n$, and is the $m \times p$ matrix $\mathbf{C} = [c_{jk}]$ with entries

$$c_{jk} = a_{j1}b_{1k} + a_{j2}b_{2k} + \cdots + a_{jn}b_{nk} \qquad \text{(row } j \text{ of } \mathbf{A} \text{ times column } k \text{ of } \mathbf{B}). \tag{1}$$

This multiplication is motivated by the composition of **linear transformations** (Secs. 7.2, 7.9). It is associative, but is ***not commutative:*** if $\mathbf{AB}$ is defined, $\mathbf{BA}$ may not be defined, but even if $\mathbf{BA}$ is defined, $\mathbf{AB} \neq \mathbf{BA}$ in general. Also $\mathbf{AB} = \mathbf{0}$ may not imply $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$ or $\mathbf{BA} = \mathbf{0}$ (Secs. 7.2, 7.8). Illustrations:

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = [11], \qquad \begin{bmatrix} 3 \\ 4 \end{bmatrix}\begin{bmatrix} 1 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 4 & 8 \end{bmatrix}.$$

The **transpose** $\mathbf{A}^\mathsf{T}$ of a matrix $\mathbf{A} = [a_{jk}]$ is $\mathbf{A}^\mathsf{T} = [a_{kj}]$; rows become columns and conversely (Sec. 7.2). Here, $\mathbf{A}$ need not be square. If it is and $\mathbf{A} = \mathbf{A}^\mathsf{T}$, then $\mathbf{A}$ is called **symmetric;** if $\mathbf{A} = -\mathbf{A}^\mathsf{T}$, it is called **skew-symmetric**. For a product, $(\mathbf{AB})^\mathsf{T} = \mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$ (Sec. 7.2).

A main application of matrices concerns **linear systems of equations**

$$\mathbf{Ax} = \mathbf{b} \qquad \text{(Sec. 7.3)} \tag{2}$$

(m equations in n unknowns $x_1, \cdots, x_n$; $\mathbf{A}$ and $\mathbf{b}$ given). The most important method of solution is the **Gauss elimination** (Sec. 7.3), which reduces the system to "triangular" form by *elementary row operations,* which leave the set of solutions unchanged. (Numeric aspects and variants, such as *Doolittle's* and *Cholesky's methods,* are discussed in Secs. 20.1 and 20.2.)

www.EngineeringEBooksPdf.com

Cramer's rule (Secs. 7.6, 7.7) represents the unknowns in a system (2) of n equations in n unknowns as quotients of determinants; for numeric work it is impractical. **Determinants** (Sec. 7.7) have decreased in importance, but will retain their place in eigenvalue problems, elementary geometry, etc.

The **inverse** $\mathbf{A}^{-1}$ of a square matrix satisfies $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. It exists if and only if $\det \mathbf{A} \neq 0$. It can be computed by the *Gauss–Jordan elimination* (Sec. 7.8).

The **rank** r of a matrix $\mathbf{A}$ is the maximum number of linearly independent rows or columns of $\mathbf{A}$ or, equivalently, the number of rows of the largest square submatrix of $\mathbf{A}$ with nonzero determinant (Secs. 7.4, 7.7).

The system (2) has solutions if and only if rank $\mathbf{A}$ = rank $[\mathbf{A} \quad \mathbf{b}]$, where $[\mathbf{A} \quad \mathbf{b}]$ is the **augmented matrix** (Fundamental Theorem, Sec. 7.5).

The **homogeneous system**

$$\mathbf{A}\mathbf{x} = \mathbf{0} \tag{3}$$

has solutions $\mathbf{x} \neq \mathbf{0}$ ("nontrivial solutions") if and only if rank $\mathbf{A} < n$, in the case $m = n$ equivalently if and only if $\det \mathbf{A} = \mathbf{0}$ (Secs. 7.6, 7.7).

Vector spaces, inner product spaces, and linear transformations are discussed in Sec. 7.9. See also Sec. 7.4.

www.EngineeringEBooksPdf.com

CHAPTER 8

Linear Algebra: Matrix Eigenvalue Problems

A matrix eigenvalue problem considers the vector equation

$$\mathbf{Ax} = \lambda\mathbf{x}. \tag{1}$$

Here $\mathbf{A}$ is a given square matrix, λ an unknown scalar, and $\mathbf{x}$ an unknown vector. In a matrix eigenvalue problem, the task is to determine λ's and $\mathbf{x}$'s that satisfy (1). Since $\mathbf{x} = \mathbf{0}$ is always a solution for any λ and thus not interesting, we only admit solutions with $\mathbf{x} \neq \mathbf{0}$.

The solutions to (1) are given the following names: The λ's that satisfy (1) are called **eigenvalues of A** and the corresponding nonzero $\mathbf{x}$'s that also satisfy (1) are called **eigenvectors of A**.

From this rather innocent looking vector equation flows an amazing amount of relevant theory and an incredible richness of applications. Indeed, eigenvalue problems come up all the time in engineering, physics, geometry, numerics, theoretical mathematics, biology, environmental science, urban planning, economics, psychology, and other areas. Thus, in your career you are likely to encounter eigenvalue problems.

We start with a basic and thorough introduction to eigenvalue problems in Sec. 8.1 and explain (1) with several simple matrices. This is followed by a section devoted entirely to applications ranging from mass–spring systems of physics to population control models of environmental science. We show you these diverse examples to train your skills in modeling and solving eigenvalue problems. Eigenvalue problems for real symmetric, skew-symmetric, and orthogonal matrices are discussed in Sec. 8.3 and their complex counterparts (which are important in modern physics) in Sec. 8.5. In Sec. 8.4 we show how by diagonalizing a matrix, we obtain its eigenvalues.

COMMENT. ***Numerics for eigenvalues* (Secs. 20.6–20.9) *can be studied immediately after this chapter*.**

Prerequisite: Chap. 7.
Sections that may be omitted in a shorter course: 8.4, 8.5.
References and Answers to Problems: App. 1 Part B, App. 2.

www.EngineeringEBooksPdf.com

The following chart identifies where different types of eigenvalue problems appear in the book.

Topic	Where to find it
Matrix Eigenvalue Problem (algebraic eigenvalue problem)	Chap. 8
Eigenvalue Problems in Numerics	Secs. 20.6–20.9
Eigenvalue Problem for ODEs (Sturm–Liouville problems)	Secs. 11.5, 11.6
Eigenvalue Problems for Systems of ODEs	Chap. 4
Eigenvalue Problems for PDEs	Secs. 12.3–12.11

8.1 The Matrix Eigenvalue Problem. Determining Eigenvalues and Eigenvectors

Consider multiplying nonzero vectors by a given square matrix, such as

$$\begin{bmatrix} 6 & 3 \\ 4 & 7 \end{bmatrix}\begin{bmatrix} 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 33 \\ 27 \end{bmatrix}, \qquad \begin{bmatrix} 6 & 3 \\ 4 & 7 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 30 \\ 40 \end{bmatrix}.$$

We want to see what influence the multiplication of the given matrix has on the vectors. In the first case, we get a totally new vector with a different direction and different length when compared to the original vector. This is what usually happens and is of no interest here. In the second case something interesting happens. The multiplication produces a vector $[30 \quad 40]^{\mathsf{T}} = 10\,[3 \quad 4]^{\mathsf{T}}$, which means the new vector has the same direction as the original vector. The scale constant, which we denote by λ is 10. *The problem of systematically finding such* λ's *and nonzero vectors for a given square matrix will be the theme of this chapter*. It is called the *matrix eigenvalue* problem or, more commonly, the *eigenvalue* problem.

We formalize our observation. Let $\mathbf{A} = [a_{jk}]$ be a given nonzero square matrix of dimension $n \times n$. Consider the following vector equation:

$$\mathbf{Ax} = \lambda\mathbf{x}. \tag{1}$$

The problem of finding nonzero $\mathbf{x}$'s and λ's that satisfy equation (1) is called an eigenvalue problem.

Remark. So $\mathbf{A}$ is a given square (!) matrix, $\mathbf{x}$ is an unknown vector, and λ is an unknown scalar. Our task is to find λ's and nonzero $\mathbf{x}$'s that satisfy (1). Geometrically, we are looking for vectors, $\mathbf{x}$, for which the multiplication by $\mathbf{A}$ has the same effect as the multiplication by a scalar λ; in other words, $\mathbf{Ax}$ should be proportional to $\mathbf{x}$. Thus, the multiplication has the effect of producing, from the original vector $\mathbf{x}$, a new vector $\lambda\mathbf{x}$ that has the same or opposite (minus sign) direction as the original vector. (This was all demonstrated in our intuitive opening example. Can you see that the second equation in that example satisfies (1) with $\lambda = 10$ and $\mathbf{x} = [3 \quad 4]^{\mathsf{T}}$, and $\mathbf{A}$ the given 2×2 matrix? Write it out.) Now why do we require $\mathbf{x}$ to be nonzero? The reason is that $\mathbf{x} = \mathbf{0}$ is always a solution of (1) for any value of λ, because $\mathbf{A0} = \mathbf{0}$. This is of no interest.

www.EngineeringEBooksPdf.com

We introduce more terminology. A value of λ, for which (1) has a solution $\mathbf{x} \neq \mathbf{0}$, is called an **eigenvalue** or *characteristic value* of the matrix $\mathbf{A}$. Another term for λ is a *latent* root. ("Eigen" is German and means "proper" or "characteristic."). The corresponding solutions $\mathbf{x} \neq \mathbf{0}$ of (1) are called the **eigenvectors** or *characteristic vectors* of $\mathbf{A}$ corresponding to that eigenvalue λ. The set of all the eigenvalues of $\mathbf{A}$ is called the **spectrum** of $\mathbf{A}$. We shall see that the spectrum consists of at least one eigenvalue and at most of n numerically different eigenvalues. The largest of the absolute values of the eigenvalues of $\mathbf{A}$ is called the *spectral radius* of $\mathbf{A}$, a name to be motivated later.

How to Find Eigenvalues and Eigenvectors

Now, with the new terminology for (1), we can just say that the problem of determining the eigenvalues and eigenvectors of a matrix is called an eigenvalue problem. (However, more precisely, we are considering an algebraic eigenvalue problem, as opposed to an eigenvalue problem involving an ODE or PDE, as considered in Secs. 11.5 and 12.3, or an integral equation.)

Eigenvalues have a very large number of applications in diverse fields such as in engineering, geometry, physics, mathematics, biology, environmental science, economics, psychology, and other areas. You will encounter applications for elastic membranes, Markov processes, population models, and others in this chapter.

Since, from the viewpoint of engineering applications, eigenvalue problems are the most important problems in connection with matrices, the student should carefully follow our discussion.

Example 1 demonstrates how to systematically solve a simple eigenvalue problem.

EXAMPLE 1 Determination of Eigenvalues and Eigenvectors

We illustrate all the steps in terms of the matrix

$$\mathbf{A} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix}.$$

Solution. **(a)** ***Eigenvalues.*** These must be determined ***first***. Equation (1) is

$$\mathbf{Ax} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \qquad \text{in components,} \qquad \begin{aligned} -5x_1 + 2x_2 &= \lambda x_1 \\ 2x_1 - 2x_2 &= \lambda x_2. \end{aligned}$$

Transferring the terms on the right to the left, we get

$$(2^*) \qquad \begin{aligned} (-5-\lambda)x_1 + \quad 2x_2 \quad &= 0 \\ 2x_1 \quad + (-2-\lambda)x_2 &= 0. \end{aligned}$$

This can be written in matrix notation

$$(3^*) \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

because (1) is $\mathbf{Ax} - \lambda\mathbf{x} = \mathbf{Ax} - \lambda\mathbf{Ix} = (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, which gives (3*). We see that this is a ***homogeneous*** linear system. By Cramer's theorem in Sec. 7.7 it has a nontrivial solution $\mathbf{x} \neq \mathbf{0}$ (an eigenvector of $\mathbf{A}$ we are looking for) if and only if its coefficient determinant is zero, that is,

$$(4^*) \qquad D(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -5-\lambda & 2 \\ 2 & -2-\lambda \end{vmatrix} = (-5-\lambda)(-2-\lambda) - 4 = \lambda^2 + 7\lambda + 6 = 0.$$

www.EngineeringEBooksPdf.com

We call $D(\lambda)$ the **characteristic determinant** or, if expanded, the **characteristic polynomial**, and $D(\lambda) = 0$ the **characteristic equation** of $\mathbf{A}$. The solutions of this quadratic equation are $\lambda_1 = -1$ and $\lambda_2 = -6$. These are the eigenvalues of $\mathbf{A}$.

($\mathbf{b}_1$) ***Eigenvector of* $\mathbf{A}$ *corresponding to*** λ_1. This vector is obtained from (2*) with $\lambda = \lambda_1 = -1$, that is,

$$\begin{aligned} -4x_1 + 2x_2 &= 0 \\ 2x_1 - x_2 &= 0. \end{aligned}$$

A solution is $x_2 = 2x_1$, as we see from either of the two equations, so that we need only one of them. This determines an eigenvector corresponding to $\lambda_1 = -1$ up to a scalar multiple. If we choose $x_1 = 1$, we obtain the eigenvector

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \qquad \text{Check:} \qquad \mathbf{A}\mathbf{x}_1 = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix} = (-1)\mathbf{x}_1 = \lambda_1\mathbf{x}_1.$$

($\mathbf{b}_2$) ***Eigenvector of* $\mathbf{A}$ *corresponding to*** λ_2. For $\lambda = \lambda_2 = -6$, equation (2*) becomes

$$\begin{aligned} x_1 + 2x_2 &= 0 \\ 2x_1 + 4x_2 &= 0. \end{aligned}$$

A solution is $x_2 = -x_1/2$ with arbitrary x_1. If we choose $x_1 = 2$, we get $x_2 = -1$. Thus an eigenvector of $\mathbf{A}$ corresponding to $\lambda_2 = -6$ is

$$\mathbf{x}_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \qquad \text{Check:} \qquad \mathbf{A}\mathbf{x}_2 = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} -12 \\ 6 \end{bmatrix} = (-6)\mathbf{x}_2 = \lambda_2\mathbf{x}_2.$$

For the matrix in the intuitive opening example at the start of Sec. 8.1, the characteristic equation is $\lambda^2 - 13\lambda + 30 = (\lambda - 10)(\lambda - 3) = 0$. The eigenvalues are $\{10, \ 3\}$. Corresponding eigenvectors are $[3 \quad 4]^\mathsf{T}$ and $[-1 \quad 1]^\mathsf{T}$, respectively. The reader may want to verify this. ■

This example illustrates the general case as follows. Equation (1) written in components is

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= \lambda x_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= \lambda x_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= \lambda x_n. \end{aligned}$$

Transferring the terms on the right side to the left side, we have

$$\begin{aligned} (a_{11} - \lambda)x_1 + \quad a_{12}x_2 \quad + \cdots + \quad a_{1n}x_n \quad &= 0 \\ a_{21}x_1 \quad + (a_{22} - \lambda)x_2 + \cdots + \quad a_{2n}x_n \quad &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 \quad + \quad a_{n2}x_2 \quad + \cdots + (a_{nn} - \lambda)x_n &= 0. \end{aligned} \tag{2}$$

In matrix notation,

$$\boxed{(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.} \tag{3}$$

www.EngineeringEBooksPdf.com

By Cramer's theorem in Sec. 7.7, this homogeneous linear system of equations has a nontrivial solution if and only if the corresponding determinant of the coefficients is zero:

$$D(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{4}$$

$\mathbf{A} - \lambda\mathbf{I}$ is called the **characteristic matrix** and $D(\lambda)$ the **characteristic determinant** of $\mathbf{A}$. Equation (4) is called the **characteristic equation** of $\mathbf{A}$. By developing $D(\lambda)$ we obtain a polynomial of nth degree in λ. This is called the **characteristic polynomial** of $\mathbf{A}$.

This proves the following important theorem.

THEOREM 1

Eigenvalues

The eigenvalues of a square matrix $\mathbf{A}$ *are the roots of the characteristic equation* (4) *of* $\mathbf{A}$.

Hence an $n \times n$ *matrix has at least one eigenvalue and at most* n *numerically different eigenvalues.*

For larger n, the actual computation of eigenvalues will, in general, require the use of Newton's method (Sec. 19.2) or another numeric approximation method in Secs. 20.7–20.9.

The eigen*values *must be determined first. Once these are known, corresponding eigen*vectors* are obtained from the system (2), for instance, by the Gauss elimination, where λ is the eigenvalue for which an eigenvector is wanted. This is what we did in Example 1 and shall do again in the examples below. (To prevent misunderstandings: *numeric approximation methods*, such as in Sec. 20.8, may determine eigen*vectors first*.)

Eigenvectors have the following properties.

THEOREM 2

Eigenvectors, Eigenspace

If $\mathbf{w}$ *and* $\mathbf{x}$ *are eigenvectors of a matrix* $\mathbf{A}$ *corresponding to* ***the same*** *eigenvalue* λ, *so are* $\mathbf{w} + \mathbf{x}$ (provided $\mathbf{x} \neq -\mathbf{w}$) *and* $k\mathbf{x}$ *for any* $k \neq 0$.

Hence the eigenvectors corresponding to one and the same eigenvalue λ *of* $\mathbf{A}$, *together with* $\mathbf{0}$, *form a vector space* (cf. Sec. 7.4), *called the* **eigenspace** *of* $\mathbf{A}$ *corresponding to that* λ.

PROOF $\mathbf{Aw} = \lambda\mathbf{w}$ and $\mathbf{Ax} = \lambda\mathbf{x}$ *imply* $\mathbf{A}(\mathbf{w} + \mathbf{x}) = \mathbf{Aw} + \mathbf{Ax} = \lambda\mathbf{w} + \lambda\mathbf{x} = \lambda(\mathbf{w} + \mathbf{x})$ and $\mathbf{A}(k\mathbf{w}) = k(\mathbf{Aw}) = k(\lambda\mathbf{w}) = \lambda(k\mathbf{w})$; hence $\mathbf{A}(k\mathbf{w} + \ell\mathbf{x}) = \lambda(k\mathbf{w} + \ell\mathbf{x})$. ■

In particular, *an eigenvector* $\mathbf{x}$ *is determined only up to a constant factor*. Hence we can **normalize** $\mathbf{x}$, that is, multiply it by a scalar to get a unit vector (see Sec. 7.9). For instance, $\mathbf{x}_1 = [1 \quad 2]^\mathsf{T}$ in Example 1 has the length $\|\mathbf{x}_1\| = \sqrt{1^2 + 2^2} = \sqrt{5}$; hence $[1/\sqrt{5} \quad 2/\sqrt{5}]^\mathsf{T}$ is a normalized eigenvector (a unit eigenvector).

www.EngineeringEBooksPdf.com

Examples 2 and 3 will illustrate that an $n \times n$ matrix may have n linearly independent eigenvectors, or it may have fewer than n. In Example 4 we shall see that a *real* matrix may have *complex* eigenvalues and eigenvectors.

EXAMPLE 2 Multiple Eigenvalues

Find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{bmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{bmatrix}.$$

Solution. For our matrix, the characteristic determinant gives the characteristic equation

$$-\lambda^3 - \lambda^2 + 21\lambda + 45 = 0.$$

The roots (eigenvalues of $\mathbf{A}$) are $\lambda_1 = 5$, $\lambda_2 = \lambda_3 = -3$. (If you have trouble finding roots, you may want to use a root finding algorithm such as Newton's method (Sec. 19.2). Your CAS or scientific calculator can find roots. However, to really learn and remember this material, you have to do some exercises with paper and pencil.) To find eigenvectors, we apply the Gauss elimination (Sec. 7.3) to the system $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, first with $\lambda = 5$ and then with $\lambda = -3$. For $\lambda = 5$ the characteristic matrix is

$$\mathbf{A} - \lambda\mathbf{I} = \mathbf{A} - 5\mathbf{I} = \begin{bmatrix} -7 & 2 & -3 \\ 2 & -4 & -6 \\ -1 & -2 & -5 \end{bmatrix}. \qquad \text{It row-reduces to} \qquad \begin{bmatrix} -7 & 2 & -3 \\ 0 & -\frac{24}{7} & -\frac{48}{7} \\ 0 & 0 & 0 \end{bmatrix}.$$

Hence it has rank 2. Choosing $x_3 = -1$ we have $x_2 = 2$ from $-\frac{24}{7}x_2 - \frac{48}{7}x_3 = 0$ and then $x_1 = 1$ from $-7x_1 + 2x_2 - 3x_3 = 0$. Hence an eigenvector of $\mathbf{A}$ corresponding to $\lambda = 5$ is $\mathbf{x}_1 = [1 \quad 2 \quad -1]^{\mathsf{T}}$.

For $\lambda = -3$ the characteristic matrix

$$\mathbf{A} - \lambda\mathbf{I} = \mathbf{A} + 3\mathbf{I} = \begin{bmatrix} 1 & 2 & -3 \\ 2 & 4 & -6 \\ -1 & -2 & 3 \end{bmatrix} \qquad \text{row-reduces to} \qquad \begin{bmatrix} 1 & 2 & -3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Hence it has rank 1. From $x_1 + 2x_2 - 3x_3 = 0$ we have $x_1 = -2x_2 + 3x_3$. Choosing $x_2 = 1, x_3 = 0$ and $x_2 = 0, x_3 = 1$, we obtain two linearly independent eigenvectors of $\mathbf{A}$ corresponding to $\lambda = -3$ [as they must exist by (5), Sec. 7.5, with rank $= 1$ and $n = 3$],

$$\mathbf{x}_2 = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}$$

and

$$\mathbf{x}_3 = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}.$$

■

The order M_λ of an eigenvalue λ as a root of the characteristic polynomial is called the **algebraic multiplicity** of λ. The number m_λ of linearly independent eigenvectors corresponding to λ is called the **geometric multiplicity** of λ. Thus m_λ is the dimension of the eigenspace corresponding to this λ.

www.EngineeringEBooksPdf.com

Since the characteristic polynomial has degree n, the sum of all the algebraic multiplicities must equal n. In Example 2 for $\lambda = -3$ we have $m_\lambda = M_\lambda = 2$. In general, $m_\lambda \leqq M_\lambda$, as can be shown. The difference $\Delta_\lambda = M_\lambda - m_\lambda$ is called the **defect** of λ. Thus $\Delta_{-3} = 0$ in Example 2, but positive defects Δ_λ can easily occur:

EXAMPLE 3 Algebraic Multiplicity, Geometric Multiplicity. Positive Defect

The characteristic equation of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{is} \quad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ 0 & -\lambda \end{vmatrix} = \lambda^2 = 0.$$

Hence $\lambda = 0$ is an eigenvalue of algebraic multiplicity $M_0 = 2$. But its geometric multiplicity is only $m_0 = 1$, since eigenvectors result from $-0x_1 + x_2 = 0$, hence $x_2 = 0$, in the form $[x_1 \quad 0]^\mathsf{T}$. Hence for $\lambda = 0$ the defect is $\Delta_0 = 1$.

Similarly, the characteristic equation of the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 0 & 3 \end{bmatrix} \quad \text{is} \quad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 3-\lambda & 2 \\ 0 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 = 0.$$

Hence $\lambda = 3$ is an eigenvalue of algebraic multiplicity $M_3 = 2$, but its geometric multiplicity is only $m_3 = 1$, since eigenvectors result from $0x_1 + 2x_2 = 0$ in the form $[x_1 \quad 0]^\mathsf{T}$. ■

EXAMPLE 4 Real Matrices with Complex Eigenvalues and Eigenvectors

Since real polynomials may have complex roots (which then occur in conjugate pairs), a real matrix may have complex eigenvalues and eigenvectors. For instance, the characteristic equation of the skew-symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad \text{is} \quad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ -1 & -\lambda \end{vmatrix} = \lambda^2 + 1 = 0.$$

It gives the eigenvalues $\lambda_1 = i\,(= \sqrt{-1})$, $\lambda_2 = -i$. Eigenvectors are obtained from $-ix_1 + x_2 = 0$ and $ix_1 + x_2 = 0$, respectively, and we can choose $x_1 = 1$ to get

$$\begin{bmatrix} 1 \\ i \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ -i \end{bmatrix}.$$ ■

In the next section we shall need the following simple theorem.

THEOREM 3 Eigenvalues of the Transpose

The transpose $\mathbf{A}^\mathsf{T}$ *of a square matrix* $\mathbf{A}$ *has the same eigenvalues as* $\mathbf{A}$.

PROOF Transposition does not change the value of the characteristic determinant, as follows from Theorem 2d in Sec. 7.7. ■

Having gained a first impression of matrix eigenvalue problems, we shall illustrate their importance with some typical applications in Sec. 8.2.

www.EngineeringEBooksPdf.com

PROBLEM SET 8.1

1–16 EIGENVALUES, EIGENVECTORS

Find the eigenvalues. Find the corresponding eigenvectors. Use the given λ or factor in Probs. 11 and 15.

1. $\begin{bmatrix} 3.0 & 0 \\ 0 & -0.6 \end{bmatrix}$

2. $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

3. $\begin{bmatrix} 5 & -2 \\ 9 & -6 \end{bmatrix}$

4. $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$

5. $\begin{bmatrix} 0 & 3 \\ -3 & 0 \end{bmatrix}$

6. $\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$

7. $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$

8. $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$

9. $\begin{bmatrix} 0.8 & -0.6 \\ 0.6 & 0.8 \end{bmatrix}$

10. $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

11. $\begin{bmatrix} 6 & 2 & -2 \\ 2 & 5 & 0 \\ -2 & 0 & 7 \end{bmatrix}, \quad \lambda = 3$

12. $\begin{bmatrix} 3 & 5 & 3 \\ 0 & 4 & 6 \\ 0 & 0 & 1 \end{bmatrix}$

13. $\begin{bmatrix} 13 & 5 & 2 \\ 2 & 7 & -8 \\ 5 & 4 & 7 \end{bmatrix}$

14. $\begin{bmatrix} 2 & 0 & -1 \\ 0 & \frac{1}{2} & 0 \\ 1 & 0 & 4 \end{bmatrix}$

15. $\begin{bmatrix} -1 & 0 & 12 & 0 \\ 0 & -1 & 0 & 12 \\ 0 & 0 & -1 & -4 \\ 0 & 0 & -4 & -1 \end{bmatrix}, \quad (\lambda + 1)^2$

16. $\begin{bmatrix} -3 & 0 & 4 & 2 \\ 0 & 1 & -2 & 4 \\ 2 & 4 & -1 & -2 \\ 0 & 2 & -2 & 3 \end{bmatrix}$

17–20 LINEAR TRANSFORMATIONS AND EIGENVALUES

Find the matrix $\mathbf{A}$ in the linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{x} = [x_1 \quad x_2]^\mathsf{T}$ ($\mathbf{x} = [x_1 \quad x_2 \quad x_3]^\mathsf{T}$) are Cartesian coordinates. Find the eigenvalues and eigenvectors and explain their geometric meaning.

17. Counterclockwise rotation through the angle $\pi/2$ about the origin in R^2.

18. Reflection about the x_1-axis in R^2.

19. Orthogonal projection (perpendicular projection) of R^2 onto the x_2-axis.

20. Orthogonal projection of R^3 onto the plane $x_2 = x_1$.

21–25 GENERAL PROBLEMS

21. Nonzero defect. Find further 2×2 and 3×3 matrices with positive defect. See Example 3.

22. Multiple eigenvalues. Find further 2×2 and 3×3 matrices with multiple eigenvalues. See Example 2.

23. Complex eigenvalues. Show that the eigenvalues of a real matrix are real or complex conjugate in pairs.

24. Inverse matrix. Show that $\mathbf{A}^{-1}$ exists if and only if the eigenvalues $\lambda_1, \cdots, \lambda_n$ are all nonzero, and then $\mathbf{A}^{-1}$ has the eigenvalues $1/\lambda_1, \cdots, 1/\lambda_n$.

25. Transpose. Illustrate Theorem 3 with examples of your own.

8.2 Some Applications of Eigenvalue Problems

We have selected some typical examples from the wide range of applications of matrix eigenvalue problems. The last example, that is, Example 4, shows an application involving vibrating springs and ODEs. It falls into the domain of Chapter 4, which covers matrix eigenvalue problems related to ODE's modeling mechanical systems and electrical

www.EngineeringEBooksPdf.com

networks. Example 4 is included to keep our discussion independent of Chapter 4. (However, the reader not interested in ODEs may want to skip Example 4 without loss of continuity.)

EXAMPLE 1 Stretching of an Elastic Membrane

An elastic membrane in the x_1x_2-plane with boundary circle $x_1^2 + x_2^2 = 1$ (Fig. 160) is stretched so that a point P: (x_1, x_2) goes over into the point Q: (y_1, y_2) given by

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{A}\mathbf{x} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \quad \text{in components,} \quad \begin{aligned} y_1 &= 5x_1 + 3x_2 \\ y_2 &= 3x_1 + 5x_2. \end{aligned} \tag{1}$$

Find the **principal directions**, that is, the directions of the position vector $\mathbf{x}$ of P for which the direction of the position vector $\mathbf{y}$ of Q is the same or exactly opposite. What shape does the boundary circle take under this deformation?

Solution. We are looking for vectors $\mathbf{x}$ such that $\mathbf{y} = \lambda\mathbf{x}$. Since $\mathbf{y} = \mathbf{A}\mathbf{x}$, this gives $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, the equation of an eigenvalue problem. In components, $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ is

$$\begin{aligned} 5x_1 + 3x_2 &= \lambda x_1 \\ 3x_1 + 5x_2 &= \lambda x_2 \end{aligned} \quad \text{or} \quad \begin{aligned} (5-\lambda)x_1 + \quad 3x_2 \quad &= 0 \\ 3x_1 \quad + (5-\lambda)x_2 &= 0. \end{aligned} \tag{2}$$

The characteristic equation is

$$\begin{vmatrix} 5-\lambda & 3 \\ 3 & 5-\lambda \end{vmatrix} = (5-\lambda)^2 - 9 = 0. \tag{3}$$

Its solutions are $\lambda_1 = 8$ and $\lambda_2 = 2$. These are the eigenvalues of our problem. For $\lambda = \lambda_1 = 8$, our system (2) becomes

$$\begin{aligned} -3x_1 + 3x_2 &= 0, \\ 3x_1 - 3x_2 &= 0. \end{aligned} \quad \left| \quad \begin{aligned} &\text{Solution } x_2 = x_1,\ x_1 \text{ arbitrary,} \\ &\text{for instance, } x_1 = x_2 = 1. \end{aligned} \right.$$

For $\lambda_2 = 2$, our system (2) becomes

$$\begin{aligned} 3x_1 + 3x_2 &= 0, \\ 3x_1 + 3x_2 &= 0. \end{aligned} \quad \left| \quad \begin{aligned} &\text{Solution } x_2 = -x_1,\ x_1 \text{ arbitrary,} \\ &\text{for instance, } x_1 = 1, x_2 = -1. \end{aligned} \right.$$

We thus obtain as eigenvectors of $\mathbf{A}$, for instance, $[1 \quad 1]^\mathsf{T}$ corresponding to λ_1 and $[1 \quad -1]^\mathsf{T}$ corresponding to λ_2 (or a nonzero scalar multiple of these). These vectors make 45° and 135° angles with the positive x_1-direction. They give the principal directions, the answer to our problem. The eigenvalues show that in the principal directions the membrane is stretched by factors 8 and 2, respectively; see Fig. 160.

Accordingly, if we choose the principal directions as directions of a new Cartesian u_1u_2-coordinate system, say, with the positive u_1-semi-axis in the first quadrant and the positive u_2-semi-axis in the second quadrant of the x_1x_2-system, and if we set $u_1 = r\cos\phi$, $u_2 = r\sin\phi$, then a boundary point of the unstretched circular membrane has coordinates $\cos\phi$, $\sin\phi$. Hence, after the stretch we have

$$z_1 = 8\cos\phi, \qquad z_2 = 2\sin\phi.$$

Since $\cos^2\phi + \sin^2\phi = 1$, this shows that the deformed boundary is an ellipse (Fig. 160)

$$\frac{z_1^2}{8^2} + \frac{z_2^2}{2^2} = 1. \tag{4}$$

www.EngineeringEBooksPdf.com

Fig. 160. Undeformed and deformed membrane in Example 1

EXAMPLE 2 **Eigenvalue Problems Arising from Markov Processes**

Markov processes as considered in Example 13 of Sec. 7.2 lead to eigenvalue problems if we ask for the limit state of the process in which the state vector $\mathbf{x}$ is reproduced under the multiplication by the stochastic matrix $\mathbf{A}$ governing the process, that is, $\mathbf{Ax} = \mathbf{x}$. Hence $\mathbf{A}$ should have the eigenvalue 1, and $\mathbf{x}$ should be a corresponding eigenvector. This is of practical interest because it shows the long-term tendency of the development modeled by the process.

In that example,

$$\mathbf{A} = \begin{bmatrix} 0.7 & 0.1 & 0 \\ 0.2 & 0.9 & 0.2 \\ 0.1 & 0 & 0.8 \end{bmatrix}. \qquad \text{For the transpose,} \qquad \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.9 & 0 \\ 0 & 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Hence $\mathbf{A}^\mathsf{T}$ has the eigenvalue 1, and the same is true for $\mathbf{A}$ by Theorem 3 in Sec. 8.1. An eigenvector $\mathbf{x}$ of $\mathbf{A}$ for $\lambda = 1$ is obtained from

$$\mathbf{A} - \mathbf{I} = \begin{bmatrix} -0.3 & 0.1 & 0 \\ 0.2 & -0.1 & 0.2 \\ 0.1 & 0 & -0.2 \end{bmatrix}, \qquad \text{row-reduced to} \qquad \begin{bmatrix} -\frac{3}{10} & \frac{1}{10} & 0 \\ 0 & -\frac{1}{30} & \frac{1}{5} \\ 0 & 0 & 0 \end{bmatrix}.$$

Taking $x_3 = 1$, we get $x_2 = 6$ from $-x_2/30 + x_3/5 = 0$ and then $x_1 = 2$ from $-3x_1/10 + x_2/10 = 0$. This gives $\mathbf{x} = [2 \quad 6 \quad 1]^\mathsf{T}$. It means that in the long run, the ratio Commercial:Industrial:Residential will approach 2:6:1, provided that the probabilities given by $\mathbf{A}$ remain (about) the same. (We switched to ordinary fractions to avoid rounding errors.) ■

EXAMPLE 3 **Eigenvalue Problems Arising from Population Models. Leslie Model**

The Leslie model describes age-specified population growth, as follows. Let the oldest age attained by the females in some animal population be 9 years. Divide the population into three age classes of 3 years each. Let the "*Leslie matrix*" be

$$\mathbf{L} = [l_{jk}] = \begin{bmatrix} 0 & 2.3 & 0.4 \\ 0.6 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix} \tag{5}$$

where l_{1k} is the average number of daughters born to a single female during the time she is in age class k, and $l_{j,j-1} (j = 2, 3)$ is the fraction of females in age class $j - 1$ that will survive and pass into class j. (a) What is the number of females in each class after 3, 6, 9 years if each class initially consists of 400 females? (b) For what initial distribution will the number of females in each class change by the same proportion? What is this rate of change?

www.EngineeringEBooksPdf.com

Solution. **(a)** Initially, $\mathbf{x}_{(0)}^{\mathsf{T}} = [400 \quad 400 \quad 400]$. After 3 years,

$$\mathbf{x}_{(3)} = \mathbf{L}\mathbf{x}_{(0)} = \begin{bmatrix} 0 & 2.3 & 0.4 \\ 0.6 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix} \begin{bmatrix} 400 \\ 400 \\ 400 \end{bmatrix} = \begin{bmatrix} 1080 \\ 240 \\ 120 \end{bmatrix}.$$

Similarly, after 6 years the number of females in each class is given by $\mathbf{x}_{(6)}^{\mathsf{T}} = (\mathbf{L}\mathbf{x}_{(3)})^{\mathsf{T}} = [600 \quad 648 \quad 72]$, and after 9 years we have $\mathbf{x}_{(9)}^{\mathsf{T}} = (\mathbf{L}\mathbf{x}_{(6)})^{\mathsf{T}} = [1519.2 \quad 360 \quad 194.4]$.

(b) Proportional change means that we are looking for a distribution vector $\mathbf{x}$ such that $\mathbf{L}\mathbf{x} = \lambda\mathbf{x}$, where λ is the rate of change (growth if $\lambda > 1$, decrease if $\lambda < 1$). The characteristic equation is (develop the characteristic determinant by the first column)

$$\det(\mathbf{L} - \lambda\mathbf{I}) = -\lambda^3 - 0.6(-2.3\lambda - 0.3 \cdot 0.4) = -\lambda^3 + 1.38\lambda + 0.072 = 0.$$

A positive root is found to be (for instance, by Newton's method, Sec. 19.2) $\lambda = 1.2$. A corresponding eigenvector $\mathbf{x}$ can be determined from the characteristic matrix

$$\mathbf{A} - 1.2\mathbf{I} = \begin{bmatrix} -1.2 & 2.3 & 0.4 \\ 0.6 & -1.2 & 0 \\ 0 & 0.3 & -1.2 \end{bmatrix}, \qquad \text{say,} \qquad \mathbf{x} = \begin{bmatrix} 1 \\ 0.5 \\ 0.125 \end{bmatrix}$$

where $x_3 = 0.125$ is chosen, $x_2 = 0.5$ then follows from $0.3x_2 - 1.2x_3 = 0$, and $x_1 = 1$ from $-1.2x_1 + 2.3x_2 + 0.4x_3 = 0$. To get an initial population of 1200 as before, we multiply $\mathbf{x}$ by $1200/(1 + 0.5 + 0.125) = 738$. *Answer:* Proportional growth of the numbers of females in the three classes will occur if the initial values are 738, 369, 92 in classes 1, 2, 3, respectively. The growth rate will be 1.2 per 3 years. ■

EXAMPLE 4 Vibrating System of Two Masses on Two Springs (Fig. 161)

Mass–spring systems involving several masses and springs can be treated as eigenvalue problems. For instance, the mechanical system in Fig. 161 is governed by the system of ODEs

$$\begin{aligned} y_1'' &= -3y_1 - 2(y_1 - y_2) = -5y_1 + 2y_2 \\ y_2'' &= \qquad\quad -2(y_2 - y_1) = \quad 2y_1 - 2y_2 \end{aligned} \tag{6}$$

where y_1 and y_2 are the displacements of the masses from rest, as shown in the figure, and primes denote derivatives with respect to time t. In vector form, this becomes

$$y'' = \begin{bmatrix} y_1'' \\ y_2'' \end{bmatrix} = \mathbf{A}\mathbf{y} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \tag{7}$$

Fig. 161. Masses on springs in Example 4

www.EngineeringEBooksPdf.com

We try a vector solution of the form

$$\mathbf{y} = \mathbf{x}e^{\omega t}. \tag{8}$$

This is suggested by a mechanical system of a single mass on a spring (Sec. 2.4), whose motion is given by exponential functions (and sines and cosines). Substitution into (7) gives

$$\omega^2\mathbf{x}e^{\omega t} = A\mathbf{x}e^{\omega t}.$$

Dividing by $e^{\omega t}$ and writing $\omega^2 = \lambda$, we see that our mechanical system leads to the eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \qquad \text{where } \lambda = \omega^2. \tag{9}$$

From Example 1 in Sec. 8.1 we see that $\mathbf{A}$ has the eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -6$. Consequently, $\omega = \pm\sqrt{-1} = \pm i$ and $\sqrt{-6} = \pm i\sqrt{6}$, respectively. Corresponding eigenvectors are

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}. \tag{10}$$

From (8) we thus obtain the four complex solutions [see (10), Sec. 2.2]

$$\mathbf{x}_1 e^{\pm it} = \mathbf{x}_1(\cos t \pm i \sin t),$$
$$\mathbf{x}_2 e^{\pm i\sqrt{6}t} = \mathbf{x}_2(\cos \sqrt{6}\, t \pm i \sin \sqrt{6}\, t).$$

By addition and subtraction (see Sec. 2.2) we get the four real solutions

$$\mathbf{x}_1 \cos t, \qquad \mathbf{x}_1 \sin t, \qquad \mathbf{x}_2 \cos \sqrt{6}\, t, \qquad \mathbf{x}_2 \sin \sqrt{6}\, t.$$

A general solution is obtained by taking a linear combination of these,

$$\mathbf{y} = \mathbf{x}_1(a_1 \cos t + b_1 \sin t) + \mathbf{x}_2(a_2 \cos \sqrt{6}\, t + b_2 \sin \sqrt{6}\, t)$$

with arbitrary constants a_1, b_1, a_2, b_2 (to which values can be assigned by prescribing initial displacement and initial velocity of each of the two masses). By (10), the components of $\mathbf{y}$ are

$$y_1 = a_1 \cos t + b_1 \sin t + 2a_2 \cos \sqrt{6}\, t + 2b_2 \sin \sqrt{6}\, t$$
$$y_2 = 2a_1 \cos t + 2b_1 \sin t - a_2 \cos \sqrt{6}\, t - b_2 \sin \sqrt{6}\, t.$$

These functions describe harmonic oscillations of the two masses. Physically, this had to be expected because we have neglected damping. ■

PROBLEM SET 8.2

1–6 ELASTIC DEFORMATIONS

Given $\mathbf{A}$ in a deformation $\mathbf{y} = \mathbf{A}\mathbf{x}$, find the principal directions and corresponding factors of extension or contraction. Show the details.

1. $\begin{bmatrix} 3.0 & 1.5 \\ 1.5 & 3.0 \end{bmatrix}$ **2.** $\begin{bmatrix} 2.0 & 0.4 \\ 0.4 & 2.0 \end{bmatrix}$

3. $\begin{bmatrix} 7 & \sqrt{6} \\ \sqrt{6} & 2 \end{bmatrix}$ **4.** $\begin{bmatrix} 5 & 2 \\ 2 & 13 \end{bmatrix}$

5. $\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$ **6.** $\begin{bmatrix} 1.25 & 0.75 \\ 0.75 & 1.25 \end{bmatrix}$

7–9 MARKOV PROCESSES

Find the limit state of the Markov process modeled by the given matrix. Show the details.

7. $\begin{bmatrix} 0.2 & 0.5 \\ 0.8 & 0.5 \end{bmatrix}$

8. $\begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.3 & 0.6 & 0.1 \\ 0.3 & 0.1 & 0.6 \end{bmatrix}$ **9.** $\begin{bmatrix} 0.6 & 0.1 & 0.2 \\ 0.4 & 0.1 & 0.4 \\ 0 & 0.8 & 0.4 \end{bmatrix}$

www.EngineeringEBooksPdf.com

10–12 AGE-SPECIFIC POPULATION

Find the growth rate in the Leslie model (see Example 3) with the matrix as given. Show the details.

10. $\begin{bmatrix} 0 & 9.0 & 5.0 \\ 0.4 & 0 & 0 \\ 0 & 0.4 & 0 \end{bmatrix}$ **11.** $\begin{bmatrix} 0 & 3.45 & 0.60 \\ 0.90 & 0 & 0 \\ 0 & 0.45 & 0 \end{bmatrix}$

12. $\begin{bmatrix} 0 & 3.0 & 2.0 & 2.0 \\ 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \end{bmatrix}$

13–15 LEONTIEF MODELS[1]

13. Leontief input–output model. Suppose that three industries are interrelated so that their outputs are used as inputs by themselves, according to the 3×3 **consumption matrix**

$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} 0.1 & 0.5 & 0 \\ 0.8 & 0 & 0.4 \\ 0.1 & 0.5 & 0.6 \end{bmatrix}$$

where a_{jk} is the fraction of the output of industry k consumed (purchased) by industry j. Let p_j be the price charged by industry j for its total output. A problem is to find prices so that for each industry, total expenditures equal total income. Show that this leads to $\mathbf{Ap} = \mathbf{p}$, where $\mathbf{p} = [p_1 \quad p_2 \quad p_3]^\mathsf{T}$, and find a solution $\mathbf{p}$ with nonnegative p_1, p_2, p_3.

14. Show that a consumption matrix as considered in Prob. 13 must have column sums 1 and always has the eigenvalue 1.

15. Open Leontief input–output model. If not the whole output but only a portion of it is consumed by the industries themselves, then instead of $\mathbf{Ax} = \mathbf{x}$ (as in Prob. 13), we have $\mathbf{x} - \mathbf{Ax} = \mathbf{y}$, where $\mathbf{x} = [x_1 \quad x_2 \quad x_3]^\mathsf{T}$ is produced, $\mathbf{Ax}$ is consumed by the industries, and, thus, $\mathbf{y}$ is the net production available for other consumers. Find for what production $\mathbf{x}$ a given **demand vector** $\mathbf{y} = [0.1 \quad 0.3 \quad 0.1]^\mathsf{T}$ can be achieved if the consumption matrix is

$$\mathbf{A} = \begin{bmatrix} 0.1 & 0.4 & 0.2 \\ 0.5 & 0 & 0.1 \\ 0.1 & 0.4 & 0.4 \end{bmatrix}.$$

16–20 GENERAL PROPERTIES OF EIGENVALUE PROBLEMS

Let $\mathbf{A} = [a_{jk}]$ be an $n \times n$ matrix with (not necessarily distinct) eigenvalues $\lambda_1, \cdots, \lambda_n$. Show.

16. Trace. The sum of the main diagonal entries, called the *trace* of $\mathbf{A}$, equals the sum of the eigenvalues of $\mathbf{A}$.

17. "Spectral shift." $\mathbf{A} - k\mathbf{I}$ has the eigenvalues $\lambda_1 - k, \cdots, \lambda_n - k$ and the same eigenvectors as $\mathbf{A}$.

18. Scalar multiples, powers. $k\mathbf{A}$ has the eigenvalues $k\lambda_1, \cdots, k\lambda_n$. $\mathbf{A}^m (m = 1, 2, \cdots)$ has the eigenvalues $\lambda_1^m, \cdots, \lambda_n^m$. The eigenvectors are those of $\mathbf{A}$.

19. Spectral mapping theorem. The "**polynomial matrix**"

$$p(\mathbf{A}) = k_m\mathbf{A}^m + k_{m-1}\mathbf{A}^{m-1} + \cdots + k_1\mathbf{A} + k_0\mathbf{I}$$

has the eigenvalues

$$p(\lambda_j) = k_m\lambda_j^m + k_{m-1}\lambda_j^{m-1} + \cdots + k_1\lambda_j + k_0$$

where $j = 1, \cdots, n$, and the same eigenvectors as $\mathbf{A}$.

20. Perron's theorem. A Leslie matrix $\mathbf{L}$ with positive $l_{12}, l_{13}, l_{21}, l_{32}$ has a positive eigenvalue. (This is a special case of the Perron–Frobenius theorem in Sec. 20.7, which is difficult to prove in its general form.)

8.3 Symmetric, Skew-Symmetric, and Orthogonal Matrices

We consider three classes of real square matrices that, because of their remarkable properties, occur quite frequently in applications. The first two matrices have already been mentioned in Sec. 7.2. The goal of Sec. 8.3 is to show their remarkable properties.

[1]WASSILY LEONTIEF (1906–1999). American economist at New York University. For his input–output analysis he was awarded the Nobel Prize in 1973.

www.EngineeringEBooksPdf.com

DEFINITIONS

Symmetric, Skew-Symmetric, and Orthogonal Matrices

A ***real*** square matrix $\mathbf{A} = [a_{jk}]$ is called

symmetric if transposition leaves it unchanged,

$$\mathbf{A}^{\mathsf{T}} = \mathbf{A}, \qquad \text{thus} \qquad a_{kj} = a_{jk}, \tag{1}$$

skew-symmetric if transposition gives the negative of $\mathbf{A}$,

$$\mathbf{A}^{\mathsf{T}} = -\mathbf{A}, \qquad \text{thus} \qquad a_{kj} = -a_{jk}, \tag{2}$$

orthogonal if transposition gives the inverse of $\mathbf{A}$,

$$\mathbf{A}^{\mathsf{T}} = \mathbf{A}^{-1}. \tag{3}$$

EXAMPLE 1 Symmetric, Skew-Symmetric, and Orthogonal Matrices

The matrices

$$\begin{bmatrix} -3 & 1 & 5 \\ 1 & 0 & -2 \\ 5 & -2 & 4 \end{bmatrix}, \quad \begin{bmatrix} 0 & 9 & -12 \\ -9 & 0 & 20 \\ 12 & -20 & 0 \end{bmatrix}, \quad \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \end{bmatrix}$$

are symmetric, skew-symmetric, and orthogonal, respectively, as you should verify. Every skew-symmetric matrix has all main diagonal entries zero. (Can you prove this?) ■

Any real square matrix $\mathbf{A}$ may be written as the sum of a symmetric matrix $\mathbf{R}$ and a skew-symmetric matrix $\mathbf{S}$, where

$$\mathbf{R} = \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathsf{T}}) \qquad \text{and} \qquad \mathbf{S} = \tfrac{1}{2}(\mathbf{A} - \mathbf{A}^{\mathsf{T}}). \tag{4}$$

EXAMPLE 2 Illustration of Formula (4)

$$\mathbf{A} = \begin{bmatrix} 9 & 5 & 2 \\ 2 & 3 & -8 \\ 5 & 4 & 3 \end{bmatrix} = \mathbf{R} + \mathbf{S} = \begin{bmatrix} 9.0 & 3.5 & 3.5 \\ 3.5 & 3.0 & -2.0 \\ 3.5 & -2.0 & 3.0 \end{bmatrix} + \begin{bmatrix} 0 & 1.5 & -1.5 \\ -1.5 & 0 & -6.0 \\ 1.5 & 6.0 & 0 \end{bmatrix}$$ ■

THEOREM 1 Eigenvalues of Symmetric and Skew-Symmetric Matrices

(a) *The eigenvalues of a symmetric matrix are real.*

(b) *The eigenvalues of a skew-symmetric matrix are pure imaginary or zero.*

This basic theorem (and an extension of it) will be proved in Sec. 8.5.

www.EngineeringEBooksPdf.com

EXAMPLE 3 Eigenvalues of Symmetric and Skew-Symmetric Matrices

The matrices in (1) and (7) of Sec. 8.2 are symmetric and have real eigenvalues. The skew-symmetric matrix in Example 1 has the eigenvalues 0, $-25i$, and $25i$. (Verify this.) The following matrix has the real eigenvalues 1 and 5 but is not symmetric. Does this contradict Theorem 1?

$$\begin{bmatrix} 3 & 4 \\ 1 & 3 \end{bmatrix}$$

■

Orthogonal Transformations and Orthogonal Matrices

Orthogonal transformations are transformations

$$\mathbf{y} = \mathbf{A}\mathbf{x} \qquad \text{where } \mathbf{A} \text{ is an orthogonal matrix.} \tag{5}$$

With each vector $\mathbf{x}$ in R^n such a transformation assigns a vector $\mathbf{y}$ in R^n. For instance, the plane rotation through an angle θ

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{6}$$

is an orthogonal transformation. It can be shown that any orthogonal transformation in the plane or in three-dimensional space is a **rotation** (possibly combined with a reflection in a straight line or a plane, respectively).

The main reason for the importance of orthogonal matrices is as follows.

THEOREM 2 Invariance of Inner Product

An orthogonal transformation preserves the value of the **inner product** *of vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *in* R^n, *defined by*

$$\mathbf{a} \bullet \mathbf{b} = \mathbf{a}^{\mathsf{T}}\mathbf{b} = [a_1 \ \cdots \ a_n] \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}. \tag{7}$$

That is, for any $\mathbf{a}$ *and* $\mathbf{b}$ *in* R^n, *orthogonal* $n \times n$ *matrix* $\mathbf{A}$, *and* $\mathbf{u} = \mathbf{A}\mathbf{a}, \mathbf{v} = \mathbf{A}\mathbf{b}$ *we have* $\mathbf{u} \bullet \mathbf{v} = \mathbf{a} \bullet \mathbf{b}$.

Hence the transformation also preserves the **length** *or* **norm** *of any vector* $\mathbf{a}$ *in* R^n *given by*

$$\|\mathbf{a}\| = \sqrt{\mathbf{a} \bullet \mathbf{a}} = \sqrt{\mathbf{a}^{\mathsf{T}}\mathbf{a}}. \tag{8}$$

PROOF Let $\mathbf{A}$ be orthogonal. Let $\mathbf{u} = \mathbf{A}\mathbf{a}$ and $\mathbf{v} = \mathbf{A}\mathbf{b}$. We must show that $\mathbf{u} \bullet \mathbf{v} = \mathbf{a} \bullet \mathbf{b}$. Now $(\mathbf{A}\mathbf{a})^{\mathsf{T}} = \mathbf{a}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}$ by (10d) in Sec. 7.2 and $\mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ by (3). Hence

$$\mathbf{u} \bullet \mathbf{v} = \mathbf{u}^{\mathsf{T}}\mathbf{v} = (\mathbf{A}\mathbf{a})^{\mathsf{T}}\mathbf{A}\mathbf{b} = \mathbf{a}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{A}\mathbf{b} = \mathbf{a}^{\mathsf{T}}\mathbf{I}\mathbf{b} = \mathbf{a}^{\mathsf{T}}\mathbf{b} = \mathbf{a} \bullet \mathbf{b}. \tag{9}$$

From this the invariance of $\|\mathbf{a}\|$ follows if we set $\mathbf{b} = \mathbf{a}$. ■

www.EngineeringEBooksPdf.com

Orthogonal matrices have further interesting properties as follows.

THEOREM 3 **Orthonormality of Column and Row Vectors**

A real square matrix is orthogonal if and only if its column vectors $\mathbf{a}_1, \cdots, \mathbf{a}_n$ *(and also its row vectors) form an* **orthonormal system**, *that is,*

$$\mathbf{a}_j \bullet \mathbf{a}_k = \mathbf{a}_j^\mathsf{T}\mathbf{a}_k = \begin{cases} 0 & \text{if} \quad j \neq k \\ 1 & \text{if} \quad j = k. \end{cases} \tag{10}$$

PROOF **(a)** Let $\mathbf{A}$ be orthogonal. Then $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{I}$. In terms of column vectors $\mathbf{a}_1, \cdots, \mathbf{a}_n$,

$$\mathbf{I} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{A}^\mathsf{T}\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\mathsf{T} \\ \vdots \\ \mathbf{a}_n^\mathsf{T} \end{bmatrix} [\mathbf{a}_1 \cdots \mathbf{a}_n] = \begin{bmatrix} \mathbf{a}_1^\mathsf{T}\mathbf{a}_1 & \mathbf{a}_1^\mathsf{T}\mathbf{a}_2 & \cdots & \mathbf{a}_1^\mathsf{T}\mathbf{a}_n \\ \cdot & \cdot & \cdots & \cdot \\ \mathbf{a}_n^\mathsf{T}\mathbf{a}_1 & \mathbf{a}_n^\mathsf{T}\mathbf{a}_2 & \cdots & \mathbf{a}_n^\mathsf{T}\mathbf{a}_n \end{bmatrix}. \tag{11}$$

The last equality implies (10), by the definition of the $n \times n$ unit matrix $\mathbf{I}$. From (3) it follows that the inverse of an orthogonal matrix is orthogonal (see CAS Experiment 12). Now the column vectors of $\mathbf{A}^{-1}(=\mathbf{A}^\mathsf{T})$ are the row vectors of $\mathbf{A}$. Hence the row vectors of $\mathbf{A}$ also form an orthonormal system.

(b) Conversely, if the column vectors of $\mathbf{A}$ satisfy (10), the off-diagonal entries in (11) must be 0 and the diagonal entries 1. Hence $\mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{I}$, as (11) shows. Similarly, $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{I}$. This implies $\mathbf{A}^\mathsf{T} = \mathbf{A}^{-1}$ because also $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ and the inverse is unique. Hence $\mathbf{A}$ is orthogonal. Similarly when the row vectors of $\mathbf{A}$ form an orthonormal system, by what has been said at the end of part (**a**). ■

THEOREM 4 **Determinant of an Orthogonal Matrix**

The determinant of an orthogonal matrix has the value $+1$ *or* -1.

PROOF From $\det \mathbf{AB} = \det \mathbf{A} \det \mathbf{B}$ (Sec. 7.8, Theorem 4) and $\det \mathbf{A}^\mathsf{T} = \det \mathbf{A}$ (Sec. 7.7, Theorem 2d), we get for an orthogonal matrix

$$1 = \det \mathbf{I} = \det(\mathbf{A}\mathbf{A}^{-1}) = \det(\mathbf{A}\mathbf{A}^\mathsf{T}) = \det \mathbf{A} \det \mathbf{A}^\mathsf{T} = (\det \mathbf{A})^2. \qquad ■$$

EXAMPLE 4 **Illustration of Theorems 3 and 4**

The last matrix in Example 1 and the matrix in (6) illustrate Theorems 3 and 4 because their determinants are -1 and $+1$, as you should verify. ■

THEOREM 5 **Eigenvalues of an Orthogonal Matrix**

The eigenvalues of an orthogonal matrix $\mathbf{A}$ *are real or complex conjugates in pairs and have absolute value* 1.

www.EngineeringEBooksPdf.com

PROOF The first part of the statement holds for any real matrix $\mathbf{A}$ because its characteristic polynomial has real coefficients, so that its zeros (the eigenvalues of $\mathbf{A}$) must be as indicated. The claim that $|\lambda| = 1$ will be proved in Sec. 8.5. ■

EXAMPLE 5 Eigenvalues of an Orthogonal Matrix

The orthogonal matrix in Example 1 has the characteristic equation

$$-\lambda^3 + \tfrac{2}{3}\lambda^2 + \tfrac{2}{3}\lambda - 1 = 0.$$

Now one of the eigenvalues must be real (why?), hence $+1$ or -1. Trying, we find -1. Division by $\lambda + 1$ gives $-(\lambda^2 - 5\lambda/3 + 1) = 0$ and the two eigenvalues $(5 + i\sqrt{11})/6$ and $(5 - i\sqrt{11})/6$, which have absolute value 1. Verify all of this. ■

Looking back at this section, you will find that the numerous basic results it contains have relatively short, straightforward proofs. This is typical of large portions of matrix eigenvalue theory.

PROBLEM SET 8.3

1–10 SPECTRUM

Are the following matrices symmetric, skew-symmetric, or orthogonal? Find the spectrum of each, thereby illustrating Theorems 1 and 5. Show your work in detail.

1. $\begin{bmatrix} 0.8 & 0.6 \\ -0.6 & 0.8 \end{bmatrix}$ **2.** $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$

3. $\begin{bmatrix} 2 & 8 \\ -8 & 2 \end{bmatrix}$ **4.** $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

5. $\begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & -2 \\ 0 & -2 & 5 \end{bmatrix}$ **6.** $\begin{bmatrix} a & k & k \\ k & a & k \\ k & k & a \end{bmatrix}$

7. $\begin{bmatrix} 0 & 9 & -12 \\ -9 & 0 & 20 \\ 12 & -20 & 0 \end{bmatrix}$ **8.** $\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$

9. $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix}$ **10.** $\begin{bmatrix} \frac{4}{9} & \frac{8}{9} & \frac{1}{9} \\ -\frac{7}{9} & \frac{4}{9} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{1}{9} & \frac{8}{9} \end{bmatrix}$

11. WRITING PROJECT. Section Summary. Summarize the main concepts and facts in this section, giving illustrative examples of your own.

12. CAS EXPERIMENT. Orthogonal Matrices.

(a) Products. Inverse. Prove that the product of two orthogonal matrices is orthogonal, and so is the inverse of an orthogonal matrix. What does this mean in terms of rotations?

(b) Rotation. Show that (6) is an orthogonal transformation. Verify that it satisfies Theorem 3. Find the inverse transformation.

(c) Powers. Write a program for computing powers $\mathbf{A}^m$ $(m = 1, 2, \cdots)$ of a 2×2 matrix $\mathbf{A}$ and their spectra. Apply it to the matrix in Prob. 1 (call it $\mathbf{A}$). To what rotation does $\mathbf{A}$ correspond? Do the eigenvalues of $\mathbf{A}^m$ have a limit as $m \to \infty$?

(d) Compute the eigenvalues of $(0.9\mathbf{A})^m$, where $\mathbf{A}$ is the matrix in Prob. 1. Plot them as points. What is their limit? Along what kind of curve do these points approach the limit?

(e) Find $\mathbf{A}$ such that $\mathbf{y} = \mathbf{A}\mathbf{x}$ is a counterclockwise rotation through 30° in the plane.

13–20 GENERAL PROPERTIES

13. Verification. Verify the statements in Example 1.

14. Verify the statements in Examples 3 and 4.

15. Sum. Are the eigenvalues of $\mathbf{A} + \mathbf{B}$ sums of the eigenvalues of $\mathbf{A}$ and of $\mathbf{B}$?

16. Orthogonality. Prove that eigenvectors of a symmetric matrix corresponding to different eigenvalues are orthogonal. Give examples.

17. Skew-symmetric matrix. Show that the inverse of a skew-symmetric matrix is skew-symmetric.

18. Do there exist nonsingular skew-symmetric $n \times n$ matrices with odd n?

19. Orthogonal matrix. Do there exist skew-symmetric orthogonal 3×3 matrices?

20. Symmetric matrix. Do there exist nondiagonal symmetric 3×3 matrices that are orthogonal?

www.EngineeringEBooksPdf.com

8.4 Eigenbases. Diagonalization. Quadratic Forms

So far we have emphasized properties of eigen*values*. We now turn to general properties of eigen*vectors*. Eigenvectors of an $n \times n$ matrix $\mathbf{A}$ may (or may not!) form a basis for R^n. If we are interested in a transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$, such an "**eigenbasis**" (basis of eigenvectors)—if it exists—is of great advantage because then we can represent any $\mathbf{x}$ in R^n uniquely as a linear combination of the eigenvectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$, say,

$$\mathbf{x} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n.$$

And, denoting the corresponding (not necessarily distinct) eigenvalues of the matrix $\mathbf{A}$ by $\lambda_1, \cdots, \lambda_n$, we have $\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_j$, so that we simply obtain

$$\begin{aligned} \mathbf{y} = \mathbf{A}\mathbf{x} &= \mathbf{A}(c_1\mathbf{x}_1 + \cdots + c_n\mathbf{x}_n) \\ &= c_1\mathbf{A}\mathbf{x}_1 + \cdots + c_n\mathbf{A}\mathbf{x}_n \\ &= c_1\lambda_1\mathbf{x}_1 + \cdots + c_n\lambda_n\mathbf{x}_n. \end{aligned} \tag{1}$$

This shows that we have decomposed the complicated action of $\mathbf{A}$ on an arbitrary vector $\mathbf{x}$ into a sum of simple actions (multiplication by scalars) on the eigenvectors of $\mathbf{A}$. This is the point of an eigenbasis.

Now if the n eigenvalues are all different, we do obtain a basis:

THEOREM 1

Basis of Eigenvectors

*If an $n \times n$ matrix $\mathbf{A}$ has n **distinct** eigenvalues, then $\mathbf{A}$ has a basis of eigenvectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$ for R^n.*

PROOF All we have to show is that $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are linearly independent. Suppose they are not. Let r be the largest integer such that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is a linearly independent set. Then $r < n$ and the set $\{\mathbf{x}_1, \cdots, \mathbf{x}_r, \mathbf{x}_{r+1}\}$ is linearly dependent. Thus there are scalars $c_1, \cdots, c_{r+1}$, not all zero, such that

$$c_1\mathbf{x}_1 + \cdots + c_{r+1}\mathbf{x}_{r+1} = \mathbf{0} \tag{2}$$

(see Sec. 7.4). Multiplying both sides by $\mathbf{A}$ and using $\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_j$, we obtain

$$\mathbf{A}(c_1\mathbf{x}_1 + \cdots + c_{r+1}\mathbf{x}_{r+1}) = c_1\lambda_1\mathbf{x}_1 + \cdots + c_{r+1}\lambda_{r+1}\mathbf{x}_{r+1} = \mathbf{A}\mathbf{0} = \mathbf{0}. \tag{3}$$

To get rid of the last term, we subtract λ_{r+1} times (2) from this, obtaining

$$c_1(\lambda_1 - \lambda_{r+1})\mathbf{x}_1 + \cdots + c_r(\lambda_r - \lambda_{r+1})\mathbf{x}_r = \mathbf{0}.$$

Here $c_1(\lambda_1 - \lambda_{r+1}) = 0, \cdots, c_r(\lambda_r - \lambda_{r+1}) = 0$ since $\{x_1, \cdots, x_r\}$ is linearly independent. Hence $c_1 = \cdots = c_r = 0$, since all the eigenvalues are distinct. But with this, (2) reduces to $c_{r+1}\mathbf{x}_{r+1} = \mathbf{0}$, hence $c_{r+1} = 0$, since $\mathbf{x}_{r+1} \neq \mathbf{0}$ (an eigenvector!). This contradicts the fact that not all scalars in (2) are zero. Hence the conclusion of the theorem must hold. ■

www.EngineeringEBooksPdf.com

EXAMPLE 1 **Eigenbasis. Nondistinct Eigenvalues. Nonexistence**

The matrix $\mathbf{A} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$ has a basis of eigenvectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ corresponding to the eigenvalues $\lambda_1 = 8$, $\lambda_2 = 2$. (See Example 1 in Sec. 8.2.)

Even if not all n eigenvalues are different, a matrix $\mathbf{A}$ may still provide an eigenbasis for R^n. See Example 2 in Sec. 8.1, where $n = 3$.

On the other hand, $\mathbf{A}$ ***may not have enough linearly independent eigenvectors to make up a basis***. For instance, $\mathbf{A}$ in Example 3 of Sec. 8.1 is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and has only one eigenvector} \quad \begin{bmatrix} k \\ 0 \end{bmatrix} \quad (k \neq 0, \text{ arbitrary}).$$

Actually, eigenbases exist under much more general conditions than those in Theorem 1. An important case is the following.

THEOREM 2 **Symmetric Matrices**

A symmetric matrix has an orthonormal basis of eigenvectors for R^n.

For a proof (which is involved) see Ref. [B3], vol. 1, pp. 270–272.

EXAMPLE 2 **Orthonormal Basis of Eigenvectors**

The first matrix in Example 1 is symmetric, and an orthonormal basis of eigenvectors is $[1/\sqrt{2} \quad 1/\sqrt{2}]^\mathsf{T}$, $[1/\sqrt{2} \quad -1/\sqrt{2}]^\mathsf{T}$.

Similarity of Matrices. Diagonalization

Eigenbases also play a role in reducing a matrix $\mathbf{A}$ to a diagonal matrix whose entries are the eigenvalues of $\mathbf{A}$. This is done by a "similarity transformation," which is defined as follows (and will have various applications in numerics in Chap. 20).

DEFINITION **Similar Matrices. Similarity Transformation**

An $n \times n$ matrix $\hat{\mathbf{A}}$ is called **similar** to an $n \times n$ matrix $\mathbf{A}$ if

$$\hat{\mathbf{A}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \tag{4}$$

for some (nonsingular!) $n \times n$ matrix $\mathbf{P}$. This transformation, which gives $\hat{\mathbf{A}}$ from $\mathbf{A}$, is called a **similarity transformation**.

The key property of this transformation is that it preserves the eigenvalues of $\mathbf{A}$:

THEOREM 3 **Eigenvalues and Eigenvectors of Similar Matrices**

If $\hat{\mathbf{A}}$ is similar to $\mathbf{A}$, then $\hat{\mathbf{A}}$ has the same eigenvalues as $\mathbf{A}$.

Furthermore, if $\mathbf{x}$ is an eigenvector of $\mathbf{A}$, then $\mathbf{y} = \mathbf{P}^{-1}\mathbf{x}$ is an eigenvector of $\hat{\mathbf{A}}$ corresponding to the same eigenvalue.

www.EngineeringEBooksPdf.com

PROOF From $\mathbf{Ax} = \lambda\mathbf{x}$ (λ an eigenvalue, $\mathbf{x} \neq \mathbf{0}$) we get $\mathbf{P}^{-1}\mathbf{Ax} = \lambda\mathbf{P}^{-1}\mathbf{x}$. Now $\mathbf{I} = \mathbf{PP}^{-1}$. By this ***identity trick*** the equation $\mathbf{P}^{-1}\mathbf{Ax} = \lambda\mathbf{P}^{-1}\mathbf{x}$ gives

$$\mathbf{P}^{-1}\mathbf{Ax} = \mathbf{P}^{-1}\mathbf{AIx} = \mathbf{P}^{-1}\mathbf{APP}^{-1}\mathbf{x} = (\mathbf{P}^{-1}\mathbf{AP})\mathbf{P}^{-1}\mathbf{x} = \hat{\mathbf{A}}(\mathbf{P}^{-1}\mathbf{x}) = \lambda\mathbf{P}^{-1}\mathbf{x}.$$

Hence λ is an eigenvalue of $\hat{\mathbf{A}}$ and $\mathbf{P}^{-1}\mathbf{x}$ a corresponding eigenvector. Indeed, $\mathbf{P}^{-1}\mathbf{x} \neq \mathbf{0}$ because $\mathbf{P}^{-1}\mathbf{x} = \mathbf{0}$ would give $\mathbf{x} = \mathbf{Ix} = \mathbf{PP}^{-1}\mathbf{x} = \mathbf{P0} = \mathbf{0}$, contradicting $\mathbf{x} \neq \mathbf{0}$. ■

EXAMPLE 3 Eigenvalues and Vectors of Similar Matrices

Let,
$$\mathbf{A} = \begin{bmatrix} 6 & -3 \\ 4 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{P} = \begin{bmatrix} 1 & 3 \\ 1 & 4 \end{bmatrix}.$$

Then
$$\hat{\mathbf{A}} = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 6 & -3 \\ 4 & -1 \end{bmatrix}\begin{bmatrix} 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}.$$

Here $\mathbf{P}^{-1}$ was obtained from (4*) in Sec. 7.8 with $\det \mathbf{P} = 1$. We see that $\hat{\mathbf{A}}$ has the eigenvalues $\lambda_1 = 3$, $\lambda_2 = 2$. The characteristic equation of $\mathbf{A}$ is $(6 - \lambda)(-1 - \lambda) + 12 = \lambda^2 - 5\lambda + 6 = 0$. It has the roots (the eigenvalues of $\mathbf{A}$) $\lambda_1 = 3$, $\lambda_2 = 2$, confirming the first part of Theorem 3.

We confirm the second part. From the first component of $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ we have $(6 - \lambda)x_1 - 3x_2 = 0$. For $\lambda = 3$ this gives $3x_1 - 3x_2 = 0$, say, $\mathbf{x}_1 = [1 \quad 1]^\mathsf{T}$. For $\lambda = 2$ it gives $4x_1 - 3x_2 = 0$, say, $\mathbf{x}_2 = [3 \quad 4]^\mathsf{T}$. In Theorem 3 we thus have

$$\mathbf{y}_1 = \mathbf{P}^{-1}\mathbf{x}_1 = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{y}_2 = \mathbf{P}^{-1}\mathbf{x}_2 = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Indeed, these are eigenvectors of the diagonal matrix $\hat{\mathbf{A}}$.

Perhaps we see that $\mathbf{x}_1$ and $\mathbf{x}_2$ are the columns of $\mathbf{P}$. This suggests the general method of transforming a matrix $\mathbf{A}$ to diagonal form $\mathbf{D}$ by using $\mathbf{P} = \mathbf{X}$, the matrix with eigenvectors as columns. ■

By a suitable similarity transformation we can now transform a matrix $\mathbf{A}$ to a diagonal matrix $\mathbf{D}$ whose diagonal entries are the eigenvalues of $\mathbf{A}$:

THEOREM 4 Diagonalization of a Matrix

If an $n \times n$ matrix $\mathbf{A}$ has a basis of eigenvectors, then

$$\mathbf{D} = \mathbf{X}^{-1}\mathbf{AX} \tag{5}$$

is diagonal, with the eigenvalues of $\mathbf{A}$ as the entries on the main diagonal. Here $\mathbf{X}$ is the matrix with these eigenvectors as column vectors. Also,

$$\mathbf{D}^m = \mathbf{X}^{-1}\mathbf{A}^m\mathbf{X} \qquad (m = 2, 3, \cdots). \tag{5*}$$

www.EngineeringEBooksPdf.com

PROOF Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be a basis of eigenvectors of $\mathbf{A}$ for R^n. Let the corresponding eigenvalues of $\mathbf{A}$ be $\lambda_1, \cdots, \lambda_n$, respectively, so that $\mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1, \cdots, \mathbf{A}\mathbf{x}_n = \lambda_n\mathbf{x}_n$. Then $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$ has rank n, by Theorem 3 in Sec. 7.4. Hence $\mathbf{X}^{-1}$ exists by Theorem 1 in Sec. 7.8. We claim that

$$\mathbf{A}\mathbf{X} = \mathbf{A}[\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_n] = [\mathbf{A}\mathbf{x}_1 \quad \cdots \quad \mathbf{A}\mathbf{x}_n] = [\lambda_1\mathbf{x}_1 \quad \cdots \quad \lambda_n\mathbf{x}_n] = \mathbf{X}\mathbf{D} \tag{6}$$

where $\mathbf{D}$ is the diagonal matrix as in (5). The fourth equality in (6) follows by direct calculation. (Try it for $n = 2$ and then for general n.) The third equality uses $\mathbf{A}\mathbf{x}_k = \lambda_k\mathbf{x}_k$. The second equality results if we note that the first column of $\mathbf{A}\mathbf{X}$ is $\mathbf{A}$ times the first column of $\mathbf{X}$, which is $\mathbf{x}_1$, and so on. For instance, when $n = 2$ and we write $\mathbf{x}_1 = [x_{11} \quad x_{21}]$, $\mathbf{x}_2 = [x_{12} \quad x_{22}]$, we have

$$\begin{aligned}\mathbf{A}\mathbf{X} = \mathbf{A}[\mathbf{x}_1 \quad \mathbf{x}_2] &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11}x_{11} + a_{12}x_{21} & a_{11}x_{12} + a_{12}x_{22} \\ a_{21}x_{11} + a_{22}x_{21} & a_{21}x_{12} + a_{22}x_{22} \end{bmatrix} = [\mathbf{A}\mathbf{x}_1 \quad \mathbf{A}\mathbf{x}_2].\end{aligned}$$

Column 1 Column 2

If we multiply (6) by $\mathbf{X}^{-1}$ from the left, we obtain (5). Since (5) is a similarity transformation, Theorem 3 implies that $\mathbf{D}$ has the same eigenvalues as $\mathbf{A}$. Equation (5*) follows if we note that

$$\mathbf{D}^2 = \mathbf{D}\mathbf{D} = (\mathbf{X}^{-1}\mathbf{A}\mathbf{X})(\mathbf{X}^{-1}\mathbf{A}\mathbf{X}) = \mathbf{X}^{-1}\mathbf{A}(\mathbf{X}\mathbf{X}^{-1})\mathbf{A}\mathbf{X} = \mathbf{X}^{-1}\mathbf{A}\mathbf{A}\mathbf{X} = \mathbf{X}^{-1}\mathbf{A}^2\mathbf{X}, \quad \text{etc.}$$ ■

EXAMPLE 4 Diagonalization

Diagonalize

$$\mathbf{A} = \begin{bmatrix} 7.3 & 0.2 & -3.7 \\ -11.5 & 1.0 & 5.5 \\ 17.7 & 1.8 & -9.3 \end{bmatrix}.$$

Solution. The characteristic determinant gives the characteristic equation $-\lambda^3 - \lambda^2 + 12\lambda = 0$. The roots (eigenvalues of $\mathbf{A}$) are $\lambda_1 = 3$, $\lambda_2 = -4$, $\lambda_3 = 0$. By the Gauss elimination applied to $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ with $\lambda = \lambda_1, \lambda_2, \lambda_3$ we find eigenvectors and then $\mathbf{X}^{-1}$ by the Gauss–Jordan elimination (Sec. 7.8, Example 1). The results are

$$\begin{bmatrix} -1 \\ 3 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix}, \qquad \mathbf{X}^{-1} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix}.$$

Calculating $\mathbf{A}\mathbf{X}$ and multiplying by $\mathbf{X}^{-1}$ from the left, we thus obtain

$$\mathbf{D} = \mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix} \begin{bmatrix} -3 & -4 & 0 \\ 9 & 4 & 0 \\ -3 & -12 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$ ■

www.EngineeringEBooksPdf.com

Quadratic Forms. Transformation to Principal Axes

By definition, a **quadratic form** Q in the components $x_1, \cdots, x_n$ of a vector $\mathbf{x}$ is a sum of n^2 terms, namely,

$$
\begin{aligned}
Q = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} &= \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_jx_k \\
&= a_{11}x_1^2 + a_{12}x_1x_2 + \cdots + a_{1n}x_1x_n \\
&\quad + a_{21}x_2x_1 + a_{22}x_2^2 + \cdots + a_{2n}x_2x_n \\
&\quad + \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\quad + a_{n1}x_nx_1 + a_{n2}x_nx_2 + \cdots + a_{nn}x_n^2.
\end{aligned} \tag{7}
$$

$\mathbf{A} = [a_{jk}]$ is called the **coefficient matrix** of the form. We may assume that $\mathbf{A}$ is ***symmetric***, because we can take off-diagonal terms together in pairs and write the result as a sum of two equal terms; see the following example.

EXAMPLE 5 Quadratic Form. Symmetric Coefficient Matrix

Let

$$\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = [x_1 \quad x_2]\begin{bmatrix} 3 & 4 \\ 6 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3x_1^2 + 4x_1x_2 + 6x_2x_1 + 2x_2^2 = 3x_1^2 + 10x_1x_2 + 2x_2^2.$$

Here $4 + 6 = 10 = 5 + 5$. From the corresponding ***symmetric*** matrix $\mathbf{C} = [c_{jk}]$, where $c_{jk} = \frac{1}{2}(a_{jk} + a_{kj})$, thus $c_{11} = 3$, $c_{12} = c_{21} = 5$, $c_{22} = 2$, we get the same result; indeed,

$$\mathbf{x}^\mathsf{T}\mathbf{C}\mathbf{x} = [x_1 \quad x_2]\begin{bmatrix} 3 & 5 \\ 5 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3x_1^2 + 5x_1x_2 + 5x_2x_1 + 2x_2^2 = 3x_1^2 + 10x_1x_2 + 2x_2^2. \qquad \blacksquare$$

Quadratic forms occur in physics and geometry, for instance, in connection with conic sections (ellipses $x_1^2/a^2 + x_2^2/b^2 = 1$, etc.) and quadratic surfaces (cones, etc.). Their transformation to principal axes is an important practical task related to the diagonalization of matrices, as follows.

By Theorem 2, the *symmetric* coefficient matrix $\mathbf{A}$ of (7) has an orthonormal basis of eigenvectors. Hence if we take these as column vectors, we obtain a matrix $\mathbf{X}$ that is orthogonal, so that $\mathbf{X}^{-1} = \mathbf{X}^\mathsf{T}$. From (5) we thus have $\mathbf{A} = \mathbf{X}\mathbf{D}\mathbf{X}^{-1} = \mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}$. Substitution into (7) gives

$$Q = \mathbf{x}^\mathsf{T}\mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}\mathbf{x}. \tag{8}$$

If we set $\mathbf{X}^\mathsf{T}\mathbf{x} = \mathbf{y}$, then, since $\mathbf{X}^\mathsf{T} = \mathbf{X}^{-1}$, we have $\mathbf{X}^{-1}\mathbf{x} = \mathbf{y}$ and thus obtain

$$\mathbf{x} = \mathbf{X}\mathbf{y}. \tag{9}$$

Furthermore, in (8) we have $\mathbf{x}^\mathsf{T}\mathbf{X} = (\mathbf{X}^\mathsf{T}\mathbf{x})^\mathsf{T} = \mathbf{y}^\mathsf{T}$ and $\mathbf{X}^\mathsf{T}\mathbf{x} = \mathbf{y}$, so that Q becomes simply

$$Q = \mathbf{y}^\mathsf{T}\mathbf{D}\mathbf{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2. \tag{10}$$

www.EngineeringEBooksPdf.com

This proves the following basic theorem.

THEOREM 5

Principal Axes Theorem

The substitution (9) *transforms a quadratic form*

$$Q = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_j x_k \qquad (a_{kj} = a_{jk})$$

to the principal axes form or **canonical form** (10), *where* $\lambda_1, \cdots, \lambda_n$ *are the* (*not necessarily distinct*) *eigenvalues of the* (*symmetric!*) *matrix* $\mathbf{A}$, *and* $\mathbf{X}$ *is an orthogonal matrix with corresponding eigenvectors* $\mathbf{x}_1, \cdots, \mathbf{x}_n$, *respectively, as column vectors.*

EXAMPLE 6 Transformation to Principal Axes. Conic Sections

Find out what type of conic section the following quadratic form represents and transform it to principal axes:

$$Q = 17x_1^2 - 30x_1x_2 + 17x_2^2 = 128.$$

Solution. We have $Q = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}$, where

$$\mathbf{A} = \begin{bmatrix} 17 & -15 \\ -15 & 17 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

This gives the characteristic equation $(17 - \lambda)^2 - 15^2 = 0$. It has the roots $\lambda_1 = 2, \lambda_2 = 32$. Hence (10) becomes

$$Q = 2y_1^2 + 32y_2^2.$$

We see that $Q = 128$ represents the ellipse $2y_1^2 + 32y_2^2 = 128$, that is,

$$\frac{y_1^2}{8^2} + \frac{y_2^2}{2^2} = 1.$$

If we want to know the direction of the principal axes in the x_1x_2-coordinates, we have to determine normalized eigenvectors from $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ with $\lambda = \lambda_1 = 2$ and $\lambda = \lambda_2 = 32$ and then use (9). We get

$$\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix},$$

hence

$$\mathbf{x} = \mathbf{X}\mathbf{y} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \qquad \begin{aligned} x_1 &= y_1/\sqrt{2} - y_2/\sqrt{2} \\ x_2 &= y_1/\sqrt{2} + y_2/\sqrt{2}. \end{aligned}$$

This is a 45° rotation. Our results agree with those in Sec. 8.2, Example 1, except for the notations. See also Fig. 160 in that example. ■

www.EngineeringEBooksPdf.com

PROBLEM SET 8.4

1–5 SIMILAR MATRICES HAVE EQUAL EIGENVALUES

Verify this for $\mathbf{A}$ and $\mathbf{A} = \mathbf{P}^{-1}\mathbf{AP}$. If $\mathbf{y}$ is an eigenvector of $\mathbf{P}$, show that $\mathbf{x} = \mathbf{Py}$ are eigenvectors of $\mathbf{A}$. Show the details of your work.

1. $\mathbf{A} = \begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} -4 & 2 \\ 3 & -1 \end{bmatrix}$

2. $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 7 & -5 \\ 10 & -7 \end{bmatrix}$

3. $\mathbf{A} = \begin{bmatrix} 8 & -4 \\ 2 & 2 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 0.28 & 0.96 \\ -0.96 & 0.28 \end{bmatrix}$

4. $\mathbf{A} = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 3 & 2 \\ 1 & 0 & 1 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 2 & 0 & 3 \\ 0 & 1 & 0 \\ 3 & 0 & 5 \end{bmatrix},$
$\lambda_1 = 3$

5. $\mathbf{A} = \begin{bmatrix} -5 & 0 & 15 \\ 3 & 4 & -9 \\ -5 & 0 & 15 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

6. PROJECT. Similarity of Matrices. Similarity is basic, for instance, in designing numeric methods.

(a) Trace. By definition, the **trace** of an $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the sum of the diagonal entries,

$$\text{trace } \mathbf{A} = a_{11} + a_{22} + \cdots + a_{nn}.$$

Show that the trace equals the sum of the eigenvalues, each counted as often as its algebraic multiplicity indicates. Illustrate this with the matrices $\mathbf{A}$ in Probs. 1, 3, and 5.

(b) Trace of product. Let $\mathbf{B} = [b_{jk}]$ be $n \times n$. Show that similar matrices have equal traces, by first proving

$$\text{trace } \mathbf{AB} = \sum_{i=1}^{n} \sum_{l=1}^{n} a_{il} b_{li} = \text{trace } \mathbf{BA}.$$

(c) Find a relationship between $\hat{\mathbf{A}}$ in (4) and $\hat{\mathbf{A}} = \mathbf{PAP}^{-1}$.

(d) Diagonalization. What can you do in (5) if you want to change the order of the eigenvalues in $\mathbf{D}$, for instance, interchange $d_{11} = \lambda_1$ and $d_{22} = \lambda_2$?

7. No basis. Find further 2×2 and 3×3 matrices without eigenbasis.

8. Orthonormal basis. Illustrate Theorem 2 with further examples.

9–16 DIAGONALIZATION OF MATRICES

Find an eigenbasis (a basis of eigenvectors) and diagonalize. Show the details.

9. $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ **10.** $\begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}$

11. $\begin{bmatrix} -19 & 7 \\ -42 & 16 \end{bmatrix}$ **12.** $\begin{bmatrix} -4.3 & 7.7 \\ 1.3 & 9.3 \end{bmatrix}$

13. $\begin{bmatrix} 4 & 0 & 0 \\ 12 & -2 & 0 \\ 21 & -6 & 1 \end{bmatrix}$

14. $\begin{bmatrix} -5 & -6 & 6 \\ -9 & -8 & 12 \\ -12 & -12 & 16 \end{bmatrix}, \quad \lambda_1 = -2$

15. $\begin{bmatrix} 4 & 3 & 3 \\ 3 & 6 & 1 \\ 3 & 1 & 6 \end{bmatrix}, \quad \lambda_1 = 10$

16. $\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & -4 \end{bmatrix}$

17–23 PRINCIPAL AXES. CONIC SECTIONS

What kind of conic section (or pair of straight lines) is given by the quadratic form? Transform it to principal axes. Express $\mathbf{x}^\mathsf{T} = [x_1 \quad x_2]$ in terms of the new coordinate vector $\mathbf{y}^\mathsf{T} = [y_1 \quad y_2]$, as in Example 6.

17. $7x_1^2 + 6x_1x_2 + 7x_2^2 = 200$

18. $3x_1^2 + 8x_1x_2 - 3x_2^2 = 10$

19. $3x_1^2 + 22x_1x_2 + 3x_2^2 = 0$

20. $9x_1^2 + 6x_1x_2 + x_2^2 = 10$

21. $x_1^2 - 12x_1x_2 + x_2^2 = 70$

22. $4x_1^2 + 12x_1x_2 + 13x_2^2 = 16$

23. $-11x_1^2 + 84x_1x_2 + 24x_2^2 = 156$

www.EngineeringEBooksPdf.com

24. Definiteness. A quadratic form $Q(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ and its (symmetric!) matrix $\mathbf{A}$ are called **(a) positive definite** if $Q(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$, **(b) negative definite** if $Q(\mathbf{x}) < 0$ for all $\mathbf{x} \neq \mathbf{0}$, **(c) indefinite** if $Q(\mathbf{x})$ takes both positive and negative values. (See Fig. 162.) [$Q(\mathbf{x})$ and $\mathbf{A}$ are called *positive semidefinite* (*negative semidefinite*) if $Q(\mathbf{x}) \geqq 0$ ($Q(\mathbf{x}) \leqq 0$) for all $\mathbf{x}$.] Show that a necessary and sufficient condition for (a), (b), and (c) is that the eigenvalues of $\mathbf{A}$ are (a) all positive, (b) all negative, and (c) both positive and negative.

Hint. Use Theorem 5.

25. Definiteness. A necessary and sufficient condition for positive definiteness of a quadratic form $Q(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ with *symmetric* matrix $\mathbf{A}$ is that all the **principal minors** are positive (see Ref. [B3], vol. 1, p. 306), that is,

$$a_{11} > 0, \qquad \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix} > 0,$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} > 0, \qquad \cdots, \qquad \det \mathbf{A} > 0.$$

Show that the form in Prob. 22 is positive definite, whereas that in Prob. 23 is indefinite.

(*a*) Positive definite form

(*b*) Negative definite form

(*c*) Indefinite form

Fig. 162. Quadratic forms in two variables (Problem 24)

8.5 Complex Matrices and Forms. *Optional*

The three classes of matrices in Sec. 8.3 have complex counterparts which are of practical interest in certain applications, for instance, in quantum mechanics. This is mainly because of their spectra as shown in Theorem 1 in this section. The second topic is about extending quadratic forms of Sec. 8.4 to complex numbers. (The reader who wants to brush up on complex numbers may want to consult Sec. 13.1.)

Notations

$\bar{\mathbf{A}} = [\bar{a}_{jk}]$ is obtained from $\mathbf{A} = [a_{jk}]$ by replacing each entry $a_{jk} = \alpha + i\beta$ (α, β real) with its complex conjugate $\bar{a}_{jk} = \alpha - i\beta$. Also, $\bar{\mathbf{A}}^\mathsf{T} = [\bar{a}_{kj}]$ is the transpose of $\bar{\mathbf{A}}$, hence the conjugate transpose of $\mathbf{A}$.

EXAMPLE 1 Notations

$$\text{If } \mathbf{A} = \begin{bmatrix} 3+4i & 1-i \\ 6 & 2-5i \end{bmatrix}, \quad \text{then} \quad \bar{\mathbf{A}} = \begin{bmatrix} 3-4i & 1+i \\ 6 & 2+5i \end{bmatrix} \quad \text{and} \quad \bar{\mathbf{A}}^\mathsf{T} = \begin{bmatrix} 3-4i & 6 \\ 1+i & 2+5i \end{bmatrix}. \quad \blacksquare$$

www.EngineeringEBooksPdf.com

DEFINITION

Hermitian, Skew-Hermitian, and Unitary Matrices

A square matrix $\mathbf{A} = [a_{kj}]$ is called

Hermitian	if	$\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}$,	that is,	$\bar{a}_{kj} = a_{jk}$
skew-Hermitian	if	$\overline{\mathbf{A}}^{\mathsf{T}} = -\mathbf{A}$,	that is,	$\bar{a}_{kj} = -a_{jk}$
unitary	if	$\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$.		

The first two classes are named after Hermite (see footnote 13 in Problem Set 5.8).

From the definitions we see the following. If $\mathbf{A}$ is Hermitian, the entries on the main diagonal must satisfy $\bar{a}_{jj} = a_{jj}$; that is, they are real. Similarly, if $\mathbf{A}$ is skew-Hermitian, then $\bar{a}_{jj} = -a_{jj}$. If we set $a_{jj} = \alpha + i\beta$, this becomes $\alpha - i\beta = -(\alpha + i\beta)$. Hence $\alpha = 0$, so that a_{jj} must be pure imaginary or 0.

EXAMPLE 2 Hermitian, Skew-Hermitian, and Unitary Matrices

$$\mathbf{A} = \begin{bmatrix} 4 & 1 - 3i \\ 1 + 3i & 7 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 3i & 2 + i \\ -2 + i & -i \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} \frac{1}{2}i & \frac{1}{2}\sqrt{3} \\ \frac{1}{2}\sqrt{3} & \frac{1}{2}i \end{bmatrix}$$

are Hermitian, skew-Hermitian, and unitary matrices, respectively, as you may verify by using the definitions. ■

If a Hermitian matrix is real, then $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} = \mathbf{A}$. Hence a real Hermitian matrix is a symmetric matrix (Sec. 8.3).

Similarly, if a skew-Hermitian matrix is real, then $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} = -\mathbf{A}$. Hence a real skew-Hermitian matrix is a skew-symmetric matrix.

Finally, if a unitary matrix is real, then $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} = \mathbf{A}^{-1}$. Hence a real unitary matrix is an orthogonal matrix.

This shows that *Hermitian, skew-Hermitian, and unitary matrices generalize symmetric, skew-symmetric, and orthogonal matrices, respectively.*

Eigenvalues

It is quite remarkable that the matrices under consideration have spectra (sets of eigenvalues; see Sec. 8.1) that can be characterized in a general way as follows (see Fig. 163).

Fig. 163. Location of the eigenvalues of Hermitian, skew-Hermitian, and unitary matrices in the complex λ-plane

www.EngineeringEBooksPdf.com

THEOREM 1

Eigenvalues

(a) *The eigenvalues of a Hermitian matrix (and thus of a symmetric matrix) are real.*

(b) *The eigenvalues of a skew-Hermitian matrix (and thus of a skew-symmetric matrix) are pure imaginary or zero.*

(c) *The eigenvalues of a unitary matrix (and thus of an orthogonal matrix) have absolute value* 1.

EXAMPLE 3 Illustration of Theorem 1

For the matrices in Example 2 we find by direct calculation

Matrix		Characteristic Equation	Eigenvalues
A	Hermitian	$\lambda^2 - 11\lambda + 18 = 0$	$9, \quad 2$
B	Skew-Hermitian	$\lambda^2 - 2i\lambda + 8 = 0$	$4i, \quad -2i$
C	Unitary	$\lambda^2 - i\lambda - 1 = 0$	$\frac{1}{2}\sqrt{3} + \frac{1}{2}i, \quad -\frac{1}{2}\sqrt{3} + \frac{1}{2}i$

and $|\pm\frac{1}{2}\sqrt{3} + \frac{1}{2}i|^2 = \frac{3}{4} + \frac{1}{4} = 1$. ■

PROOF We prove Theorem 1. Let λ be an eigenvalue and $\mathbf{x}$ an eigenvector of $\mathbf{A}$. Multiply $\mathbf{Ax} = \lambda\mathbf{x}$ from the left by $\bar{\mathbf{x}}^\mathsf{T}$, thus $\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax} = \lambda\bar{\mathbf{x}}^\mathsf{T}\mathbf{x}$, and divide by $\bar{\mathbf{x}}^\mathsf{T}\mathbf{x} = \bar{x}_1x_1 + \cdots + \bar{\mathbf{x}}_nx_n = |x_1|^2 + \cdots + |x_n|^2$, which is real and not 0 because $\mathbf{x} \neq \mathbf{0}$. This gives

$$\lambda = \frac{\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax}}{\bar{\mathbf{x}}^\mathsf{T}\mathbf{x}}. \tag{1}$$

(a) If $\mathbf{A}$ is Hermitian, $\bar{\mathbf{A}}^\mathsf{T} = \mathbf{A}$ or $\mathbf{A}^\mathsf{T} = \bar{\mathbf{A}}$ and we show that then the numerator in (1) is real, which makes λ real. $\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax}$ is a scalar; hence taking the transpose has no effect. Thus

$$\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax} = (\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax})^\mathsf{T} = \mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\bar{\mathbf{x}} = \mathbf{x}^\mathsf{T}\bar{\mathbf{A}}\bar{\mathbf{x}} = (\overline{\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax}}). \tag{2}$$

Hence, $\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax}$ equals its complex conjugate, so that it must be real. ($a + ib = a - ib$ implies $b = 0$.)

(b) If $\mathbf{A}$ is skew-Hermitian, $\mathbf{A}^\mathsf{T} = -\bar{\mathbf{A}}$ and instead of (2) we obtain

$$\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax} = -(\overline{\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax}}) \tag{3}$$

so that $\bar{\mathbf{x}}^\mathsf{T}\mathbf{Ax}$ equals minus its complex conjugate and is pure imaginary or 0. ($a + ib = -(a - ib)$ implies $a = 0$.)

(c) Let $\mathbf{A}$ be unitary. We take $\mathbf{Ax} = \lambda\mathbf{x}$ and its conjugate transpose

$$(\overline{\mathbf{Ax}})^\mathsf{T} = (\bar{\lambda}\bar{\mathbf{x}})^\mathsf{T} = \bar{\lambda}\bar{\mathbf{x}}^\mathsf{T}$$

and multiply the two left sides and the two right sides,

$$(\overline{\mathbf{Ax}})^\mathsf{T}\mathbf{Ax} = \bar{\lambda}\lambda\bar{\mathbf{x}}^\mathsf{T}\mathbf{x} = |\lambda|^2\bar{\mathbf{x}}^\mathsf{T}\mathbf{x}.$$

www.EngineeringEBooksPdf.com

But $\mathbf{A}$ is unitary, $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$, so that on the left we obtain

$$(\overline{\mathbf{A}}\overline{\mathbf{x}})^{\mathsf{T}}\mathbf{A}\mathbf{x} = \overline{\mathbf{x}}^{\mathsf{T}}\overline{\mathbf{A}}^{\mathsf{T}}\mathbf{A}\mathbf{x} = \overline{\mathbf{x}}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \overline{\mathbf{x}}^{\mathsf{T}}\mathbf{I}\mathbf{x} = \overline{\mathbf{x}}^{\mathsf{T}}\mathbf{x}.$$

Together, $\overline{\mathbf{x}}^{\mathsf{T}}\mathbf{x} = |\lambda|^2\overline{\mathbf{x}}^{\mathsf{T}}\mathbf{x}$. We now divide by $\overline{\mathbf{x}}^{\mathsf{T}}\mathbf{x}$ $(\neq 0)$ to get $|\lambda|^2 = 1$. Hence $|\lambda| = 1$.

This proves Theorem 1 as well as Theorems 1 and 5 in Sec. 8.3. ■

Key properties of orthogonal matrices (invariance of the inner product, orthonormality of rows and columns; see Sec. 8.3) generalize to unitary matrices in a remarkable way.

To see this, instead of R^n we now use the **complex vector space** C^n of all complex vectors with n complex numbers as components, and complex numbers as scalars. For such complex vectors the **inner product** is defined by (note the overbar for the complex conjugate)

$$\mathbf{a} \bullet \mathbf{b} = \overline{\mathbf{a}}^{\mathsf{T}}\mathbf{b}. \tag{4}$$

The **length** or **norm** of such a complex vector is a *real* number defined by

$$\|\mathbf{a}\| = \sqrt{\mathbf{a} \bullet \mathbf{a}} = \sqrt{\overline{\mathbf{a}}_j^{\mathsf{T}}\mathbf{a}} = \sqrt{\overline{a}_1 a_1 + \cdots + \overline{a}_n a_n} = \sqrt{|a_1|^2 + \cdots + |a_n|^2}. \tag{5}$$

THEOREM 2

Invariance of Inner Product

A **unitary transformation**, *that is,* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *with a unitary matrix* $\mathbf{A}$, *preserves the value of the inner product* (4), *hence also the norm* (5).

PROOF The proof is the same as that of Theorem 2 in Sec. 8.3, which the theorem generalizes. In the analog of (9), Sec. 8.3, we now have bars,

$$\mathbf{u} \bullet \mathbf{v} = \overline{\mathbf{u}}^{\mathsf{T}}\mathbf{v} = (\overline{\mathbf{A}}\overline{\mathbf{a}})^{\mathsf{T}}\mathbf{A}\mathbf{b} = \overline{\mathbf{a}}^{\mathsf{T}}\overline{\mathbf{A}}^{\mathsf{T}}\mathbf{A}\mathbf{b} = \overline{\mathbf{a}}^{\mathsf{T}}\mathbf{I}\mathbf{b} = \overline{\mathbf{a}}^{\mathsf{T}}\mathbf{b} = \mathbf{a} \bullet \mathbf{b}.$$

The complex analog of an orthonormal system of real vectors (see Sec. 8.3) is defined as follows.

DEFINITION

Unitary System

A *unitary system* is a set of complex vectors satisfying the relationships

$$\mathbf{a}_j \bullet \mathbf{a}_k = \overline{\mathbf{a}}_j^{\mathsf{T}}\mathbf{a}_k = \begin{cases} 0 & \text{if} \quad j \neq k \\ 1 & \text{if} \quad j = k. \end{cases} \tag{6}$$

Theorem 3 in Sec. 8.3 extends to complex as follows.

THEOREM 3

Unitary Systems of Column and Row Vectors

A complex square matrix is unitary if and only if its column vectors (*and also its row vectors*) *form a unitary system.*

www.EngineeringEBooksPdf.com

PROOF The proof is the same as that of Theorem 3 in Sec. 8.3, except for the bars required in $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$ and in (4) and (6) of the present section.

THEOREM 4 **Determinant of a Unitary Matrix**

Let $\mathbf{A}$ *be a unitary matrix. Then its determinant has absolute value one, that is,* $|\det \mathbf{A}| = 1$.

PROOF Similarly, as in Sec. 8.3, we obtain

$$1 = \det(\mathbf{A}\mathbf{A}^{-1}) = \det(\mathbf{A}\overline{\mathbf{A}}^{\mathsf{T}}) = \det \mathbf{A} \det \overline{\mathbf{A}}^{\mathsf{T}} = \det \mathbf{A} \det \overline{\mathbf{A}}$$
$$= \det \mathbf{A}\, \overline{\det \mathbf{A}} = |\det \mathbf{A}|^2.$$

Hence $|\det \mathbf{A}| = 1$ (where $\det \mathbf{A}$ may now be complex). ■

EXAMPLE 4 **Unitary Matrix Illustrating Theorems 1c and 2–4**

For the vectors $\mathbf{a}^{\mathsf{T}} = [2 \quad -i]$ and $\mathbf{b}^{\mathsf{T}} = [1+i \quad 4i]$ we get $\overline{\mathbf{a}}^{\mathsf{T}} = [2 \quad i]^{\mathsf{T}}$ and $\overline{\mathbf{a}}^{\mathsf{T}}\mathbf{b} = 2(1+i) - 4 = -2 + 2i$ and with

$$\mathbf{A} = \begin{bmatrix} 0.8i & 0.6 \\ 0.6 & 0.8i \end{bmatrix} \quad \text{also} \quad \mathbf{A}\mathbf{a} = \begin{bmatrix} i \\ 2 \end{bmatrix} \quad \text{and} \quad \mathbf{A}\mathbf{b} = \begin{bmatrix} -0.8 + 3.2i \\ -2.6 + 0.6i \end{bmatrix},$$

as one can readily verify. This gives $(\overline{\mathbf{A}}\overline{\mathbf{a}})^{\mathsf{T}}\mathbf{A}\mathbf{b} = -2 + 2i$, illustrating Theorem 2. The matrix is unitary. Its columns form a unitary system,

$$\overline{\mathbf{a}}_1^{\mathsf{T}}\mathbf{a}_1 = -0.8i \cdot 0.8i + 0.6^2 = 1, \qquad \overline{\mathbf{a}}_1^{\mathsf{T}}\mathbf{a}_2 = -0.8i \cdot 0.6 + 0.6 \cdot 0.8i = 0,$$
$$\overline{\mathbf{a}}_2^{\mathsf{T}}\mathbf{a}_2 = 0.6^2 + (-0.8i)0.8i = 1$$

and so do its rows. Also, $\det \mathbf{A} = -1$. The eigenvalues are $0.6 + 0.8i$ and $-0.6 + 0.8i$, with eigenvectors $[1 \quad 1]^{\mathsf{T}}$ and $[1 \quad -1]^{\mathsf{T}}$, respectively. ■

Theorem 2 in Sec. 8.4 on the existence of an eigenbasis extends to complex matrices as follows.

THEOREM 5 **Basis of Eigenvectors**

A Hermitian, skew-Hermitian, or unitary matrix has a basis of eigenvectors for C^n *that is a unitary system.*

For a proof see Ref. [B3], vol. 1, pp. 270–272 and p. 244 (Definition 2).

EXAMPLE 5 **Unitary Eigenbases**

The matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ in Example 2 have the following unitary systems of eigenvectors, as you should verify.

A: $\dfrac{1}{\sqrt{35}}[1-3i \quad 5]^{\mathsf{T}} \quad (\lambda = 9)$, $\qquad \dfrac{1}{\sqrt{14}}[1-3i \quad -2]^{\mathsf{T}} \quad (\lambda = 2)$

B: $\dfrac{1}{\sqrt{30}}[1-2i \quad -5]^{\mathsf{T}} \quad (\lambda = -2i)$, $\qquad \dfrac{1}{\sqrt{30}}[5 \quad 1+2i]^{\mathsf{T}} \quad (\lambda = 4i)$

C: $\dfrac{1}{\sqrt{2}}[1 \quad 1]^{\mathsf{T}} \quad (\lambda = \frac{1}{2}(i + \sqrt{3}))$, $\qquad \dfrac{1}{\sqrt{2}}[1 \quad -1]^{\mathsf{T}} \quad (\lambda = \frac{1}{2}(i - \sqrt{3}))$. ■

www.EngineeringEBooksPdf.com

Hermitian and Skew-Hermitian Forms

The concept of a quadratic form (Sec. 8.4) can be extended to complex. We call the numerator $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{A}\mathbf{x}$ in (1) a **form** in the components $x_1, \cdots, x_n$ of $\mathbf{x}$, which may now be complex. This form is again a sum of n^2 terms

$$
\begin{aligned}
\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{A}\mathbf{x} &= \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}\bar{x}_j x_k \\
&= \quad a_{11}\bar{x}_1 x_1 + \cdots + a_{1n}\bar{x}_1 x_n \\
&\quad + a_{21}\bar{x}_2 x_1 + \cdots + a_{2n}\bar{x}_2 x_n \\
&\quad + \cdots\cdots\cdots\cdots\cdots\cdots \\
&\quad + a_{n1}\bar{x}_n x_1 + \cdots + a_{nn}\bar{x}_n x_n.
\end{aligned} \tag{7}
$$

$\mathbf{A}$ is called its **coefficient matrix**. The form is called a **Hermitian** or **skew-Hermitian form** if $\mathbf{A}$ is Hermitian or skew-Hermitian, respectively. *The value of a Hermitian form is real, and that of a skew-Hermitian form is pure imaginary or zero.* This can be seen directly from (2) and (3) and accounts for the importance of these forms in physics. Note that (2) and (3) are valid for any vectors because, in the proof of (2) and (3), we did not use that $\mathbf{x}$ is an eigenvector but only that $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}$ is real and not 0.

EXAMPLE 6 Hermitian Form

For $\mathbf{A}$ in Example 2 and, say, $\mathbf{x} = [1 + i \quad 5i]^{\mathsf{T}}$ we get

$$
\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{A}\mathbf{x} = [1 - i \quad -5i]\begin{bmatrix} 4 & 1 - 3i \\ 1 + 3i & 7 \end{bmatrix}\begin{bmatrix} 1 + i \\ 5i \end{bmatrix} = [1 - i \quad -5i]\begin{bmatrix} 4(1 + i) + (1 - 3i) \cdot 5i \\ (1 + 3i)(1 + i) + 7 \cdot 5i \end{bmatrix} = 223. \quad \blacksquare
$$

Clearly, if $\mathbf{A}$ and $\mathbf{x}$ in (4) are real, then (7) reduces to a quadratic form, as discussed in the last section.

PROBLEM SET 8.5

1–6 EIGENVALUES AND VECTORS

Is the given matrix Hermitian? Skew-Hermitian? Unitary? Find its eigenvalues and eigenvectors.

1. $\begin{bmatrix} 6 & i \\ -i & 6 \end{bmatrix}$

2. $\begin{bmatrix} i & 1 + i \\ -1 + i & 0 \end{bmatrix}$

3. $\begin{bmatrix} \frac{1}{2} & i\sqrt{\frac{3}{4}} \\ i\sqrt{\frac{3}{4}} & \frac{1}{2} \end{bmatrix}$

4. $\begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$

5. $\begin{bmatrix} i & 0 & 0 \\ 0 & 0 & i \\ 0 & i & 0 \end{bmatrix}$

6. $\begin{bmatrix} 0 & 2 + 2i & 0 \\ 2 - 2i & 0 & 2 + 2i \\ 0 & 2 - 2i & 0 \end{bmatrix}$

7. Pauli spin matrices. Find the eigenvalues and eigenvectors of the so-called *Pauli spin matrices* and show that $\mathbf{S}_x\mathbf{S}_y = i\mathbf{S}_z$, $\mathbf{S}_y\mathbf{S}_x = -i\mathbf{S}_z$, $\mathbf{S}_x^2 = \mathbf{S}_y^2 = \mathbf{S}_z^2 = \mathbf{I}$, where

$$
\mathbf{S}_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{S}_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix},
$$

$$
\mathbf{S}_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.
$$

8. Eigenvectors. Find eigenvectors of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ in Examples 2 and 3.

www.EngineeringEBooksPdf.com

9–12 COMPLEX FORMS

Is the matrix **A** Hermitian or skew-Hermitian? Find $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{A}\mathbf{x}$. Show the details.

9. $\mathbf{A} = \begin{bmatrix} 4 & 3-2i \\ 3+2i & -4 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} -4i \\ 2+2i \end{bmatrix}$

10. $\mathbf{A} = \begin{bmatrix} i & -2+3i \\ 2+3i & 0 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 2i \\ 8 \end{bmatrix}$

11. $\mathbf{A} = \begin{bmatrix} i & 1 & 2+i \\ -1 & 0 & 3i \\ -2+i & 3i & i \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ i \\ -i \end{bmatrix}$

12. $\mathbf{A} = \begin{bmatrix} 1 & i & 4 \\ -i & 3 & 0 \\ 4 & 0 & 2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ i \\ -i \end{bmatrix}$

13–20 GENERAL PROBLEMS

13. Product. Show that $\overline{(\mathbf{ABC})}^{\mathsf{T}} = -\mathbf{C}^{-1}\mathbf{BA}$ for any $n \times n$ Hermitian **A**, skew-Hermitian **B**, and unitary **C**.

14. Product. Show $\overline{(\mathbf{BA})}^{\mathsf{T}} = -\mathbf{AB}$ for **A** and **B** in Example 2. For any $n \times n$ Hermitian **A** and skew-Hermitian **B**.

15. Decomposition. Show that any square matrix may be written as the sum of a Hermitian and a skew-Hermitian matrix. Give examples.

16. Unitary matrices. Prove that the product of two unitary $n \times n$ matrices and the inverse of a unitary matrix are unitary. Give examples.

17. Powers of unitary matrices in applications may sometimes be very simple. Show that $\mathbf{C}^{12} = \mathbf{I}$ in Example 2. Find further examples.

18. Normal matrix. This important concept denotes a matrix that commutes with its conjugate transpose, $\mathbf{A}\bar{\mathbf{A}}^{\mathsf{T}} = \bar{\mathbf{A}}^{\mathsf{T}}\mathbf{A}$. Prove that Hermitian, skew-Hermitian, and unitary matrices are normal. Give corresponding examples of your own.

19. Normality criterion. Prove that **A** is normal if and only if the Hermitian and skew-Hermitian matrices in Prob. 18 commute.

20. Find a simple matrix that is not normal. Find a normal matrix that is not Hermitian, skew-Hermitian, or unitary.

CHAPTER 8 REVIEW QUESTIONS AND PROBLEMS

1. In solving an eigenvalue problem, what is given and what is sought?
2. Give a few typical applications of eigenvalue problems.
3. Do there exist square matrices without eigenvalues?
4. Can a real matrix have complex eigenvalues? Can a complex matrix have real eigenvalues?
5. Does a 5×5 matrix always have a real eigenvalue?
6. What is algebraic multiplicity of an eigenvalue? Defect?
7. What is an eigenbasis? When does it exist? Why is it important?
8. When can we expect orthogonal eigenvectors?
9. State the definitions and main properties of the three classes of real matrices and of complex matrices that we have discussed.
10. What is diagonalization? Transformation to principal axes?

11–15 SPECTRUM

Find the eigenvalues. Find the eigenvectors.

11. $\begin{bmatrix} 2.5 & 0.5 \\ 0.5 & 2.5 \end{bmatrix}$ **12.** $\begin{bmatrix} -7 & 4 \\ -12 & 7 \end{bmatrix}$

13. $\begin{bmatrix} 8 & -1 \\ 5 & 2 \end{bmatrix}$

14. $\begin{bmatrix} 7 & 2 & -1 \\ 2 & 7 & 1 \\ -1 & 1 & 8.5 \end{bmatrix}$

15. $\begin{bmatrix} 0 & -3 & -6 \\ 3 & 0 & -6 \\ 6 & 6 & 0 \end{bmatrix}$

16–17 SIMILARITY

Verify that **A** and $\hat{\mathbf{A}} = \mathbf{p}^{-1}\mathbf{AP}$ have the same spectrum.

16. $\mathbf{A} = \begin{bmatrix} 19 & 12 \\ 12 & 1 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 2 & 4 \\ 4 & 2 \end{bmatrix}$

17. $\mathbf{A} = \begin{bmatrix} 7 & -4 \\ 12 & -7 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$

18. $\mathbf{A} = \begin{bmatrix} -4 & 6 & 6 \\ 0 & 2 & 0 \\ -1 & 1 & 1 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 1 & 8 & -7 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix}$

www.EngineeringEBooksPdf.com

19–21 DIAGONALIZATION

Find an eigenbasis and diagonalize.

9. $\begin{bmatrix} -1.4 & 1.0 \\ -1.0 & 1.1 \end{bmatrix}$

20. $\begin{bmatrix} 72 & -56 \\ -56 & 513 \end{bmatrix}$

21. $\begin{bmatrix} -12 & 22 & 6 \\ 8 & 2 & 6 \\ -8 & 20 & 16 \end{bmatrix}$

22–25 CONIC SECTIONS. PRINCIPAL AXES

Transform to canonical form (to principal axes). Express $[x_1 \quad x_2]^\mathsf{T}$ in terms of the new variables $[y_1 \quad y_2]^\mathsf{T}$.

22. $9x_1^2 - 6x_1x_2 + 17x_2^2 = 36$

23. $4x_1^2 + 24x_1x_2 - 14x_2^2 = 20$

24. $5x_1^2 + 24x_1x_2 - 5x_2^2 = 0$

25. $3.7x_1^2 + 3.2x_1x_2 + 1.3x_2^2 = 4.5$

SUMMARY OF CHAPTER 8
Linear Algebra: Matrix Eigenvalue Problems

The practical importance of matrix eigenvalue problems can hardly be overrated. The problems are defined by the vector equation

$$\mathbf{Ax} = \lambda\mathbf{x}. \tag{1}$$

$\mathbf{A}$ is a given square matrix. All matrices in this chapter are ***square***. λ is a scalar. To *solve* the problem (1) means to determine values of λ, called **eigenvalues** (or **characteristic values**) of $\mathbf{A}$, such that (1) has a nontrivial solution $\mathbf{x}$ (that is, $\mathbf{x} \neq \mathbf{0}$), called an **eigenvector** of $\mathbf{A}$ corresponding to that λ. An $n \times n$ matrix has at least one and at most n numerically different eigenvalues. These are the solutions of the **characteristic equation** (Sec. 8.1)

$$D(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{2}$$

$D(\lambda)$ is called the **characteristic determinant** of $\mathbf{A}$. By expanding it we get the **characteristic polynomial** of $\mathbf{A}$, which is of degree n in λ. Some typical applications are shown in Sec. 8.2.

Section 8.3 is devoted to eigenvalue problems for **symmetric** ($\mathbf{A}^\mathsf{T} = \mathbf{A}$), **skew-symmetric** ($\mathbf{A}^\mathsf{T} = -\mathbf{A}$), and **orthogonal matrices** ($\mathbf{A}^\mathsf{T} = \mathbf{A}^{-1}$). Section 8.4 concerns the diagonalization of matrices and the transformation of quadratic forms to principal axes and its relation to eigenvalues.

Section 8.5 extends Sec. 8.3 to the complex analogs of those real matrices, called **Hermitian** ($\mathbf{A}^\mathsf{T} = \mathbf{A}$), **skew-Hermitian** ($\mathbf{A}^\mathsf{T} = -\mathbf{A}$), and **unitary matrices** ($\overline{\mathbf{A}}^\mathsf{T} = \mathbf{A}^{-1}$). All the eigenvalues of a Hermitian matrix (and a symmetric one) are real. For a skew-Hermitian (and a skew-symmetric) matrix they are pure imaginary or zero. For a unitary (and an orthogonal) matrix they have absolute value 1.

www.EngineeringEBooksPdf.com

3. Linear Transformations

3.1. Linear Transformations

We shall now introduce linear transformations, the objects which we shall study in most of the remainder of this book. The reader may find it helpful to read (or reread) the discussion of functions in the Appendix, since we shall freely use the terminology of that discussion.

Definition. *Let* V *and* W *be vector spaces over the field* F. *A* **linear transformation from** V **into** W *is a function* T *from* V *into* W *such that*

$$\mathrm{T}(\mathrm{c}\alpha + \beta) = \mathrm{c}(\mathrm{T}\alpha) + \mathrm{T}\beta$$

for all α *and* β *in* V *and all scalars* c *in* F.

EXAMPLE 1. If V is any vector space, the identity transformation I, defined by $I\alpha = \alpha$, is a linear transformation from V into V. The **zero transformation** 0, defined by $0\alpha = 0$, is a linear transformation from V into V.

EXAMPLE 2. Let F be a field and let V be the space of polynomial functions f from F into F, given by

$$f(x) = c_0 + c_1x + \cdots + c_kx^k.$$

Let

$$(Df)(x) = c_1 + 2c_2x + \cdots + kc_kx^{k-1}.$$

Then D is a linear transformation from V into V—the differentiation transformation.

EXAMPLE 3. Let A be a fixed $m \times n$ matrix with entries in the field F. The function T defined by $T(X) = AX$ is a linear transformation from $F^{n\times 1}$ into $F^{m\times 1}$. The function U defined by $U(\alpha) = \alpha A$ is a linear transformation from F^m into F^n.

EXAMPLE 4. Let P be a fixed $m \times m$ matrix with entries in the field F and let Q be a fixed $n \times n$ matrix over F. Define a function T from the space $F^{m\times n}$ into itself by $T(A) = PAQ$. Then T is a linear transformation from $F^{m\times n}$ into $F^{m\times n}$, because

$$\begin{aligned} T(cA + B) &= P(cA + B)Q \\ &= (cPA + PB)Q \\ &= cPAQ + PBQ \\ &= cT(A) + T(B). \end{aligned}$$

EXAMPLE 5. Let R be the field of real numbers and let V be the space of all functions from R into R which are *continuous*. Define T by

$$(Tf)(x) = \int_0^x f(t)\, dt.$$

Then T is a linear transformation from V into V. The function Tf is not only continuous but has a continuous first derivative. The linearity of integration is one of its fundamental properties.

The reader should have no difficulty in verifying that the transformations defined in Examples 1, 2, 3, and 5 are linear transformations. We shall expand our list of examples considerably as we learn more about linear transformations.

It is important to note that if T is a linear transformation from V into W, then $T(0) = 0$; one can see this from the definition because

$$T(0) = T(0 + 0) = T(0) + T(0).$$

This point is often confusing to the person who is studying linear algebra for the first time, since he probably has been exposed to a slightly different use of the term 'linear function.' A brief comment should clear up the confusion. Suppose V is the vector space R^1. A linear transformation from V into V is then a particular type of real-valued function on the real line R. In a calculus course, one would probably call such a function linear if its graph is a straight line. A linear transformation from R^1 into R^1, according to our definition, will be a function from R into R, the graph of which is a straight line *passing through the origin*.

In addition to the property $T(0) = 0$, let us point out another property of the general linear transformation T. Such a transformation 'preserves' linear combinations; that is, if $\alpha_1, \ldots, \alpha_n$ are vectors in V and $c_1, \ldots, c_n$ are scalars, then

$$T(c_1\alpha_1 + \cdots + c_n\alpha_n) = c_1(T\alpha_1) + \cdots + c_n(T\alpha_n).$$

This follows readily from the definition. For example,

$$\begin{aligned} T(c_1\alpha_1 + c_2\alpha_2) &= c_1(T\alpha_1) + T(c_2\alpha_2) \\ &= c_1(T\alpha_1) + c_2(T\alpha_2). \end{aligned}$$

Theorem 1. *Let* V *be a finite-dimensional vector space over the field* F *and let* $\{\alpha_1, \ldots, \alpha_n\}$ *be an ordered basis for* V. *Let* W *be a vector space over the same field* F *and let* $\beta_1, \ldots, \beta_n$ *be any vectors in* W. *Then there is precisely one linear transformation* T *from* V *into* W *such that*

$$\mathrm{T}\alpha_j = \beta_j, \qquad j = 1, \ldots, n.$$

Proof. To prove there is some linear transformation T with $T\alpha_j = \beta_j$ we proceed as follows. Given α in V, there is a unique n-tuple $(x_1, \ldots, x_n)$ such that

$$\alpha = x_1\alpha_1 + \cdots + x_n\alpha_n.$$

For this vector α we define

$$T\alpha = x_1\beta_1 + \cdots + x_n\beta_n.$$

Then T is a well-defined rule for associating with each vector α in V a vector $T\alpha$ in W. From the definition it is clear that $T\alpha_j = \beta_j$ for each j. To see that T is linear, let

$$\beta = y_1\alpha_1 + \cdots + y_n\alpha_n$$

be in V and let c be any scalar. Now

$$c\alpha + \beta = (cx_1 + y_1)\alpha_1 + \cdots + (cx_n + y_n)\alpha_n$$

and so by definition

$$T(c\alpha + \beta) = (cx_1 + y_1)\beta_1 + \cdots + (cx_n + y_n)\beta_n.$$

On the other hand,

$$\begin{aligned} c(T\alpha) + T\beta &= c\sum_{i=1}^{n} x_i\beta_i + \sum_{i=1}^{n} y_i\beta_i \\ &= \sum_{i=1}^{n} (cx_i + y_i)\beta_i \end{aligned}$$

and thus

$$T(c\alpha + \beta) = c(T\alpha) + T\beta.$$

If U is a linear transformation from V into W with $U\alpha_j = \beta_j$, $j = 1, \ldots, n$, then for the vector $\alpha = \sum_{i=1}^{n} x_i\alpha_i$ we have

$$\begin{aligned} U\alpha &= U\left(\sum_{i=1}^{n} x_i\alpha_i\right) \\ &= \sum_{i=1}^{n} x_i(U\alpha_i) \\ &= \sum_{i=1}^{n} x_i\beta_i \end{aligned}$$

so that U is exactly the rule T which we defined above. This shows that the linear transformation T with $T\alpha_j = \beta_j$ is unique. ∎

Theorem 1 is quite elementary; however, it is so basic that we have stated it formally. The concept of function is very general. If V and W are (non-zero) vector spaces, there is a multitude of functions from V into W. Theorem 1 helps to underscore the fact that the functions which are linear are extremely special.

EXAMPLE 6. The vectors

$$\alpha_1 = (1, 2)$$
$$\alpha_2 = (3, 4)$$

are linearly independent and therefore form a basis for R^2. According to Theorem 1, there is a unique linear transformation from R^2 into R^3 such that

$$T\alpha_1 = (3, 2, 1)$$
$$T\alpha_2 = (6, 5, 4).$$

If so, we must be able to find $T(\epsilon_1)$. We find scalars c_1, c_2 such that $\epsilon_1 = c_1\alpha_1 + c_2\alpha_2$ and then we know that $T\epsilon_1 = c_1T\alpha_1 + c_2T\alpha_2$. If $(1, 0) = c_1(1, 2) + c_2(3, 4)$ then $c_1 = -2$ and $c_2 = 1$. Thus

$$\begin{aligned} T(1, 0) &= -2(3, 2, 1) + (6, 5, 4) \\ &= (0, 1, 2). \end{aligned}$$

EXAMPLE 7. Let T be a linear transformation from the m-tuple space F^m into the n-tuple space F^n. Theorem 1 tells us that T is uniquely determined by the sequence of vectors $\beta_1, \ldots, \beta_m$ where

$$\beta_i = T\epsilon_i, \qquad i = 1, \ldots, m.$$

In short, T is uniquely determined by the images of the standard basis vectors. The determination is

$$\alpha = (x_1, \ldots, x_m)$$
$$T\alpha = x_1\beta_1 + \cdots + x_m\beta_m.$$

If B is the $m \times n$ matrix which has row vectors $\beta_1, \ldots, \beta_m$, this says that

$$T\alpha = \alpha B.$$

In other words, if $\beta_i = (B_{i1}, \ldots, B_{in})$, then

$$T(x_1, \ldots, x_n) = [x_1 \cdots x_m] \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & & \vdots \\ B_{m1} & \cdots & B_{mn} \end{bmatrix}.$$

This is a very explicit description of the linear transformation. In Section 3.4 we shall make a serious study of the relationship between linear trans-

formations and matrices. We shall not pursue the particular description $T\alpha = \alpha B$ because it has the matrix B on the right of the vector α, and that can lead to some confusion. The point of this example is to show that we can give an explicit and reasonably simple description of all linear transformations from F^m into F^n.

If T is a linear transformation from V into W, then the range of T is not only a subset of W; it is a subspace of W. Let R_T be the range of T, that is, the set of all vectors β in W such that $\beta = T\alpha$ for some α in V. Let β_1 and β_2 be in R_T and let c be a scalar. There are vectors α_1 and α_2 in V such that $T\alpha_1 = \beta_1$ and $T\alpha_2 = \beta_2$. Since T is linear

$$\begin{aligned} T(c\alpha_1 + \alpha_2) &= cT\alpha_1 + T\alpha_2 \\ &= c\beta_1 + \beta_2, \end{aligned}$$

which shows that $c\beta_1 + \beta_2$ is also in R_T.

Another interesting subspace associated with the linear transformation T is the set N consisting of the vectors α in V such that $T\alpha = 0$. It is a subspace of V because

(a) $T(0) = 0$, so that N is non-empty;
(b) if $T\alpha_1 = T\alpha_2 = 0$, then

$$\begin{aligned} T(c\alpha_1 + \alpha_2) &= cT\alpha_1 + T\alpha_2 \\ &= c0 + 0 \\ &= 0 \end{aligned}$$

so that $c\alpha_1 + \alpha_2$ is in N.

Definition. *Let* V *and* W *be vector spaces over the field* F *and let* T *be a linear transformation from* V *into* W. *The* **null space** *of* T *is the set of all vectors* α *in* V *such that* $T\alpha = 0$.

If V *is finite-dimensional, the* **rank** *of* T *is the dimension of the range of* T *and the* **nullity** *of* T *is the dimension of the null space of* T.

The following is one of the most important results in linear algebra.

Theorem 2. *Let* V *and* W *be vector spaces over the field* F *and let* T *be a linear transformation from* V *into* W. *Suppose that* V *is finite-dimensional. Then*

$$\textit{rank}\ (\mathrm{T}) + \textit{nullity}\ (\mathrm{T}) = \textit{dim}\ \mathrm{V}.$$

Proof. Let $\{\alpha_1, \ldots, \alpha_k\}$ be a basis for N, the null space of T. There are vectors $\alpha_{k+1}, \ldots, \alpha_n$ in V such that $\{\alpha_1, \ldots, \alpha_n\}$ is a basis for V. We shall now prove that $\{T\alpha_{k+1}, \ldots, T\alpha_n\}$ is a basis for the range of T. The vectors $T\alpha_1, \ldots, T\alpha_n$ certainly span the range of T, and since $T\alpha_j = 0$, for $j \leq k$, we see that $T\alpha_{k+1}, \ldots, T\alpha_n$ span the range. To see that these vectors are independent, suppose we have scalars c_i such that

$$\sum_{i=k+1}^{n} c_i(T\alpha_i) = 0.$$

This says that

$$T\left(\sum_{i=k+1}^{n} c_i\alpha_i\right) = 0$$

and accordingly the vector $\alpha = \sum_{i=k+1}^{n} c_i\alpha_i$ is in the null space of T. Since $\alpha_1, \ldots, \alpha_k$ form a basis for N, there must be scalars $b_1, \ldots, b_k$ such that

$$\alpha = \sum_{i=1}^{k} b_i\alpha_i.$$

Thus

$$\sum_{i=1}^{k} b_i\alpha_i - \sum_{j=k+1}^{n} c_j\alpha_j = 0$$

and since $\alpha_1, \ldots, \alpha_n$ are linearly independent we must have

$$b_1 = \cdots = b_k = c_{k+1} = \cdots = c_n = 0.$$

If r is the rank of T, the fact that $T\alpha_{k+1}, \ldots, T\alpha_n$ form a basis for the range of T tells us that $r = n - k$. Since k is the nullity of T and n is the dimension of V, we are done. ∎

Theorem 3. *If* A *is an* m $\times$ n *matrix with entries in the field* F, *then*

$$\textit{row rank } (\mathrm{A}) = \textit{column rank } (\mathrm{A}).$$

Proof. Let T be the linear transformation from $F^{n\times 1}$ into $F^{m\times 1}$ defined by $T(X) = AX$. The null space of T is the solution space for the system $AX = 0$, i.e., the set of all column matrices X such that $AX = 0$. The range of T is the set of all $m \times 1$ column matrices Y such that $AX = Y$ has a solution for X. If $A_1, \ldots, A_n$ are the columns of A, then

$$AX = x_1A_1 + \cdots + x_nA_n$$

so that the range of T is the subspace spanned by the columns of A. In other words, the range of T is the column space of A. Therefore,

$$\text{rank } (T) = \text{column rank } (A).$$

Theorem 2 tells us that if S is the solution space for the system $AX = 0$, then

$$\dim S + \text{column rank } (A) = n.$$

We now refer to Example 15 of Chapter 2. Our deliberations there showed that, if r is the dimension of the row space of A, then the solution space S has a basis consisting of $n - r$ vectors:

$$\dim S = n - \text{row rank } (A).$$

It is now apparent that

$$\text{row rank } (A) = \text{column rank } (A). \quad ∎$$

The proof of Theorem 3 which we have just given depends upon

explicit calculations concerning systems of linear equations. There is a more conceptual proof which does not rely on such calculations. We shall give such a proof in Section 3.7.

Exercises

1. Which of the following functions T from R^2 into R^2 are linear transformations?

(a) $T(x_1, x_2) = (1 + x_1, x_2)$;
(b) $T(x_1, x_2) = (x_2, x_1)$;
(c) $T(x_1, x_2) = (x_1^2, x_2)$;
(d) $T(x_1, x_2) = (\sin x_1, x_2)$;
(e) $T(x_1, x_2) = (x_1 - x_2, 0)$.

2. Find the range, rank, null space, and nullity for the zero transformation and the identity transformation on a finite-dimensional space V.

3. Describe the range and the null space for the differentiation transformation of Example 2. Do the same for the integration transformation of Example 5.

4. Is there a linear transformation T from R^3 into R^2 such that $T(1, -1, 1) = (1, 0)$ and $T(1, 1, 1) = (0, 1)$?

5. If

$$\begin{aligned} \alpha_1 &= (1, -1), & \beta_1 &= (1, 0) \\ \alpha_2 &= (2, -1), & \beta_2 &= (0, 1) \\ \alpha_3 &= (-3, 2), & \beta_3 &= (1, 1) \end{aligned}$$

is there a linear transformation T from R^2 into R^2 such that $T\alpha_i = \beta_i$ for $i = 1, 2$ and 3?

6. Describe explicitly (as in Exercises 1 and 2) the linear transformation T from F^2 into F^2 such that $T\epsilon_1 = (a, b)$, $T\epsilon_2 = (c, d)$.

7. Let F be a subfield of the complex numbers and let T be the function from F^3 into F^3 defined by

$$T(x_1, x_2, x_3) = (x_1 - x_2 + 2x_3, 2x_1 + x_2, -x_1 - 2x_2 + 2x_3).$$

(a) Verify that T is a linear transformation.

(b) If (a, b, c) is a vector in F^3, what are the conditions on a, b, and c that the vector be in the range of T? What is the rank of T?

(c) What are the conditions on a, b, and c that (a, b, c) be in the null space of T? What is the nullity of T?

8. Describe explicitly a linear transformation from R^3 into R^3 which has as its range the subspace spanned by $(1, 0, -1)$ and $(1, 2, 2)$.

9. Let V be the vector space of all $n \times n$ matrices over the field F, and let B be a fixed $n \times n$ matrix. If

$$T(A) = AB - BA$$

verify that T is a linear transformation from V into V.

10. Let V be the set of all complex numbers regarded as a vector space over the

field of *real* numbers (usual operations). Find a function from V into V which is a linear transformation on the above vector space, but which is not a linear transformation on C^1, i.e., which is not complex linear.

11. Let V be the space of $n \times 1$ matrices over F and let W be the space of $m \times 1$ matrices over F. Let A be a fixed $m \times n$ matrix over F and let T be the linear transformation from V into W defined by $T(X) = AX$. Prove that T is the zero transformation if and only if A is the zero matrix.

12. Let V be an n-dimensional vector space over the field F and let T be a linear transformation from V into V such that the range and null space of T are identical. Prove that n is even. (Can you give an example of such a linear transformation T?)

13. Let V be a vector space and T a linear transformation from V into V. Prove that the following two statements about T are equivalent.

(a) The intersection of the range of T and the null space of T is the zero subspace of V.

(b) If $T(T\alpha) = 0$, then $T\alpha = 0$.

3.2. The Algebra of Linear Transformations

In the study of linear transformations from V into W, it is of fundamental importance that the set of these transformations inherits a natural vector space structure. The set of linear transformations from a space V into itself has even more algebraic structure, because ordinary composition of functions provides a 'multiplication' of such transformations. We shall explore these ideas in this section.

Theorem 4. *Let* V *and* W *be vector spaces over the field* F. *Let* T *and* U *be linear transformations from* V *into* W. *The function* (T + U) *defined by*

$$(\mathrm{T} + \mathrm{U})(\alpha) = \mathrm{T}\alpha + \mathrm{U}\alpha$$

is a linear transformation from V *into* W. *If* c *is any element of* F, *the function* (cT) *defined by*

$$(\mathrm{cT})(\alpha) = \mathrm{c}(\mathrm{T}\alpha)$$

is a linear transformation from V *into* W. *The set of all linear transformations from* V *into* W, *together with the addition and scalar multiplication defined above, is a vector space over the field* F.

Proof. Suppose T and U are linear transformations from V into W and that we define $(T + U)$ as above. Then

$$\begin{aligned}(T + U)(c\alpha + \beta) &= T(c\alpha + \beta) + U(c\alpha + \beta)\\ &= c(T\alpha) + T\beta + c(U\alpha) + U\beta\\ &= c(T\alpha + U\alpha) + (T\beta + U\beta)\\ &= c(T + U)(\alpha) + (T + U)(\beta)\end{aligned}$$

which shows that $(T + U)$ is a linear transformation. Similarly,

$$\begin{aligned}(cT)(d\alpha + \beta) &= c[T(d\alpha + \beta)] \\ &= c[d(T\alpha) + T\beta] \\ &= cd(T\alpha) + c(T\beta) \\ &= d[c(T\alpha)] + c(T\beta) \\ &= d[(cT)\alpha] + (cT)\beta\end{aligned}$$

which shows that (cT) is a linear transformation.

To verify that the set of linear transformations of V into W (together with these operations) is a vector space, one must directly check each of the conditions on the vector addition and scalar multiplication. We leave the bulk of this to the reader, and content ourselves with this comment: The zero vector in this space will be the zero transformation, which sends every vector of V into the zero vector in W; each of the properties of the two operations follows from the corresponding property of the operations in the space W. ∎

We should perhaps mention another way of looking at this theorem. If one defines sum and scalar multiple as we did above, then the set of *all* functions from V into W becomes a vector space over the field F. This has nothing to do with the fact that V is a vector space, only that V is a non-empty set. When V is a vector space we can define a linear transformation from V into W, and Theorem 4 says that the linear transformations are a subspace of the space of all functions from V into W.

We shall denote the space of linear transformations from V into W by $L(V, W)$. We remind the reader that $L(V, W)$ is defined only when V and W are vector spaces over the same field.

Theorem 5. *Let* V *be an* n*-dimensional vector space over the field* F, *and let* W *be an* m*-dimensional vector space over* F. *Then the space* L(V, W) *is finite-dimensional and has dimension* mn.

Proof. Let

$$\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\} \quad \text{and} \quad \mathcal{B}' = \{\beta_1, \ldots, \beta_m\}$$

be ordered bases for V and W, respectively. For each pair of integers (p, q) with $1 \leq p \leq m$ and $1 \leq q \leq n$, we define a linear transformation $E^{p,q}$ from V into W by

$$\begin{aligned}E^{p,q}(\alpha_i) &= \begin{cases} 0, & \text{if } i \neq q \\ \beta_p, & \text{if } i = q \end{cases} \\ &= \delta_{iq}\beta_p.\end{aligned}$$

According to Theorem 1, there is a unique linear transformation from V into W satisfying these conditions. The claim is that the mn transformations $E^{p,q}$ form a basis for $L(V, W)$.

Let T be a linear transformation from V into W. For each j, $1 \leq j \leq n$,

let $A_{ij}, \ldots, A_{mj}$ be the coordinates of the vector $T\alpha_j$ in the ordered basis $\mathcal{B}'$, i.e.,

$$T\alpha_j = \sum_{p=1}^{m} A_{pj}\beta_p. \tag{3-1}$$

We wish to show that

$$T = \sum_{p=1}^{m} \sum_{q=1}^{n} A_{pq}E^{p,q}. \tag{3-2}$$

Let U be the linear transformation in the right-hand member of (3-2). Then for each j

$$\begin{aligned} U\alpha_j &= \sum_p \sum_q A_{pq}E^{p,q}(\alpha_j) \\ &= \sum_p \sum_q A_{pq}\delta_{jq}\beta_p \\ &= \sum_{p=1}^{m} A_{pj}\beta_p \\ &= T\alpha_j \end{aligned}$$

and consequently $U = T$. Now (3-2) shows that the $E^{p,q}$ span $L(V, W)$; we must prove that they are independent. But this is clear from what we did above; for, if the transformation

$$U = \sum_p \sum_q A_{pq}E^{p,q}$$

is the zero transformation, then $U\alpha_j = 0$ for each j, so

$$\sum_{p=1}^{m} A_{pj}\beta_p = 0$$

and the independence of the β_p implies that $A_{pj} = 0$ for every p and j. ∎

Theorem 6. *Let* V, W, *and* Z *be vector spaces over the field* F. *Let* T *be a linear transformation from* V *into* W *and* U *a linear transformation from* W *into* Z. *Then the composed function* UT *defined by* $(\mathrm{UT})(\alpha) = \mathrm{U}(\mathrm{T}(\alpha))$ *is a linear transformation from* V *into* Z.

Proof.

$$\begin{aligned} (UT)(c\alpha + \beta) &= U[T(c\alpha + \beta)] \\ &= U(cT\alpha + T\beta) \\ &= c[U(T\alpha)] + U(T\beta) \\ &= c(UT)(\alpha) + (UT)(\beta). \end{aligned}$$ ∎

In what follows, we shall be primarily concerned with linear transformation of a vector space into itself. Since we would so often have to write 'T is a linear transformation from V into V,' we shall replace this with 'T is a linear operator on V.'

Definition. *If* V *is a vector space over the field* F, *a* **linear operator on** V *is a linear transformation from* V *into* V.

In the case of Theorem 6 when $V = W = Z$, so that U and T are linear operators on the space V, we see that the composition UT is again a linear operator on V. Thus the space $L(V, V)$ has a 'multiplication' defined on it by composition. In this case the operator TU is also defined, and one should note that in general $UT \neq TU$, i.e., $UT - TU \neq 0$. We should take special note of the fact that if T is a linear operator on V then we can compose T with T. We shall use the notation $T^2 = TT$, and in general $T^n = T \cdots T$ (n times) for $n = 1, 2, 3, \ldots$. We define $T^0 = I$ if $T \neq 0$.

Lemma. *Let* V *be a vector space over the field* F; *let* U, T_1 *and* T_2 *be linear operators on* V; *let* c *be an element of* F.

(a) $IU = UI = U$;
(b) $U(T_1 + T_2) = UT_1 + UT_2$; $(T_1 + T_2)U = T_1U + T_2U$;
(c) $c(UT_1) = (cU)T_1 = U(cT_1)$.

Proof. (a) This property of the identity function is obvious. We have stated it here merely for emphasis.

(b)
$$\begin{aligned}[U(T_1 + T_2)](\alpha) &= U[(T_1 + T_2)(\alpha)] \\ &= U(T_1\alpha + T_2\alpha) \\ &= U(T_1\alpha) + U(T_2\alpha) \\ &= (UT_1)(\alpha) + (UT_2)(\alpha)\end{aligned}$$

so that $U(T_1 + T_2) = UT_1 + UT_2$. Also

$$\begin{aligned}[(T_1 + T_2)U](\alpha) &= (T_1 + T_2)(U\alpha) \\ &= T_1(U\alpha) + T_2(U\alpha) \\ &= (T_1U)(\alpha) + (T_2U)(\alpha)\end{aligned}$$

so that $(T_1 + T_2)U = T_1U + T_2U$. (The reader may note that the proofs of these two distributive laws do not use the fact that T_1 and T_2 are linear, and the proof of the second one does not use the fact that U is linear either.)

(c) We leave the proof of part (c) to the reader. ∎

The contents of this lemma and a portion of Theorem 5 tell us that the vector space $L(V, V)$, together with the composition operation, is what is known as a linear algebra with identity. We shall discuss this in Chapter 4.

EXAMPLE 8. If A is an $m \times n$ matrix with entries in F, we have the linear transformation T defined by $T(X) = AX$, from $F^{n\times 1}$ into $F^{m\times 1}$. If B is a $p \times m$ matrix, we have the linear transformation U from $F^{m\times 1}$ into $F^{p\times 1}$ defined by $U(Y) = BY$. The composition UT is easily described:

$$\begin{aligned}(UT)(X) &= U(T(X)) \\ &= U(AX) \\ &= B(AX) \\ &= (BA)X.\end{aligned}$$

Thus UT is 'left multiplication by the product matrix BA.'

EXAMPLE 9. Let F be a field and V the vector space of all polynomial functions from F into F. Let D be the differentiation operator defined in Example 2, and let T be the linear operator 'multiplication by x':

$$(Tf)(x) = xf(x).$$

Then $DT \neq TD$. In fact, the reader should find it easy to verify that $DT - TD = I$, the identity operator.

Even though the 'multiplication' we have on $L(V, V)$ is not commutative, it is nicely related to the vector space operations of $L(V, V)$.

EXAMPLE 10. Let $\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}$ be an ordered basis for a vector space V. Consider the linear operators $E^{p,q}$ which arose in the proof of Theorem 5:

$$E^{p,q}(\alpha_i) = \delta_{iq}\alpha_p.$$

These n^2 linear operators form a basis for the space of linear operators on V. What is $E^{p,q}E^{r,s}$? We have

$$\begin{aligned}(E^{p,q}E^{r,s})(\alpha_i) &= E^{p,q}(\delta_{is}\alpha_r)\\ &= \delta_{is}E^{p,q}(\alpha_r)\\ &= \delta_{is}\delta_{rq}\alpha_p.\end{aligned}$$

Therefore,

$$E^{p,q}E^{r,s} = \begin{cases}0, & \text{if } r \neq q\\ E^{p,s}, & \text{if } q = r.\end{cases}$$

Let T be a linear operator on V. We showed in the proof of Theorem 5 that if

$$\begin{aligned}A_j &= [T\alpha_j]_{\mathcal{B}}\\ A &= [A_1, \ldots, A_n]\end{aligned}$$

then

$$T = \sum_p \sum_q A_{pq}E^{p,q}.$$

If

$$U = \sum_r \sum_s B_{rs}E^{r,s}$$

is another linear operator on V, then the last lemma tells us that

$$\begin{aligned}TU &= (\sum_p \sum_q A_{pq}E^{p,q})(\sum_r \sum_s B_{rs}E^{r,s})\\ &= \sum_p \sum_q \sum_r \sum_s A_{pq}B_{rs}E^{p,q}E^{r,s}.\end{aligned}$$

As we have noted, the only terms which survive in this huge sum are the terms where $q = r$, and since $E^{p,r}E^{r,s} = E^{p,s}$, we have

$$\begin{aligned}TU &= \sum_p \sum_s (\sum_r A_{pr}B_{rs})E^{p,s}\\ &= \sum_p \sum_s (AB)_{ps}E^{p,s}.\end{aligned}$$

Thus, the effect of composing T and U is to multiply the matrices A and B.

In our discussion of algebraic operations with linear transformations we have not yet said anything about invertibility. One specific question of interest is this. For which linear operators T on the space V does there exist a linear operator T^{-1} such that $TT^{-1} = T^{-1}T = I$?

The function T from V into W is called **invertible** if there exists a function U from W into V such that UT is the identity function on V and TU is the identity function on W. If T is invertible, the function U is unique and is denoted by T^{-1}. (See Appendix.) Furthermore, T is invertible if and only if

1. T is 1:1, that is, $T\alpha = T\beta$ implies $\alpha = \beta$;
2. T is onto, that is, the range of T is (all of) W.

Theorem 7. *Let* V *and* W *be vector spaces over the field* F *and let* T *be a linear transformation from* V *into* W. *If* T *is invertible, then the inverse function* T^{-1} *is a linear transformation from* W *onto* V.

Proof. We repeat ourselves in order to underscore a point. When T is one-one and onto, there is a uniquely determined inverse function T^{-1} which maps W onto V such that $T^{-1}T$ is the identity function on V, and TT^{-1} is the identity function on W. What we are proving here is that if a linear function T is invertible, then the inverse T^{-1} is also linear.

Let β_1 and β_2 be vectors in W and let c be a scalar. We wish to show that

$$T^{-1}(c\beta_1 + \beta_2) = cT^{-1}\beta_1 + T^{-1}\beta_2.$$

Let $\alpha_i = T^{-1}\beta_i$, $i = 1, 2$, that is, let α_i be the unique vector in V such that $T\alpha_i = \beta_i$. Since T is linear,

$$\begin{aligned} T(c\alpha_1 + \alpha_2) &= cT\alpha_1 + T\alpha_2 \\ &= c\beta_1 + \beta_2. \end{aligned}$$

Thus $c\alpha_1 + \alpha_2$ is the unique vector in V which is sent by T into $c\beta_1 + \beta_2$, and so

$$\begin{aligned} T^{-1}(c\beta_1 + \beta_2) &= c\alpha_1 + \alpha_2 \\ &= c(T^{-1}\beta_1) + T^{-1}\beta_2 \end{aligned}$$

and T^{-1} is linear. ∎

Suppose that we have an invertible linear transformation T from V onto W and an invertible linear transformation U from W onto Z. Then UT is invertible and $(UT)^{-1} = T^{-1}U^{-1}$. That conclusion does not require the linearity nor does it involve checking separately that UT is 1:1 and onto. All it involves is verifying that $T^{-1}U^{-1}$ is both a left and a right inverse for UT.

If T is linear, then $T(\alpha - \beta) = T\alpha - T\beta$; hence, $T\alpha = T\beta$ if and only if $T(\alpha - \beta) = 0$. This simplifies enormously the verification that T is 1:1. Let us call a linear transformation T **non-singular** if $T\gamma = 0$ implies

$\gamma = 0$, i.e., if the null space of T is $\{0\}$. Evidently, T is 1:1 if and only if T is non-singular. The extension of this remark is that non-singular linear transformations are those which preserve linear independence.

Theorem 8. *Let* T *be a linear transformation from* V *into* W. *Then* T *is non-singular if and only if* T *carries each linearly independent subset of* V *onto a linearly independent subset of* W.

Proof. First suppose that T is non-singular. Let S be a linearly independent subset of V. If $\alpha_1, \ldots, \alpha_k$ are vectors in S, then the vectors $T\alpha_1, \ldots, T\alpha_k$ are linearly independent; for if

$$c_1(T\alpha_1) + \cdots + c_k(T\alpha_k) = 0$$

then

$$T(c_1\alpha_1 + \cdots + c_k\alpha_k) = 0$$

and since T is non-singular

$$c_1\alpha_1 + \cdots + c_k\alpha_k = 0$$

from which it follows that each $c_i = 0$ because S is an independent set. This argument shows that the image of S under T is independent.

Suppose that T carries independent subsets onto independent subsets. Let α be a non-zero vector in V. Then the set S consisting of the one vector α is independent. The image of S is the set consisting of the one vector $T\alpha$, and this set is independent. Therefore $T\alpha \neq 0$, because the set consisting of the zero vector alone is dependent. This shows that the null space of T is the zero subspace, i.e., T is non-singular. ∎

EXAMPLE 11. Let F be a subfield of the complex numbers (or a field of characteristic zero) and let V be the space of polynomial functions over F. Consider the differentiation operator D and the 'multiplication by x' operator T, from Example 9. Since D sends all constants into 0, D is singular; however, V is not finite dimensional, the range of D is all of V, and it is possible to define a right inverse for D. For example, if E is the indefinite integral operator:

$$E(c_0 + c_1x + \cdots + c_nx^n) = c_0x + \frac{1}{2}c_1x^2 + \cdots + \frac{1}{n+1}c_nx^{n+1}$$

then E is a linear operator on V and $DE = I$. On the other hand, $ED \neq I$ because ED sends the constants into 0. The operator T is in what we might call the reverse situation. If $xf(x) = 0$ for all x, then $f = 0$. Thus T is non-singular and it is possible to find a left inverse for T. For example if U is the operation 'remove the constant term and divide by x':

$$U(c_0 + c_1x + \cdots + c_nx^n) = c_1 + c_2x + \cdots + c_nx^{n-1}$$

then U is a linear operator on V and $UT = I$. But $TU \neq I$ since every

function in the range of TU is in the range of T, which is the space of polynomial functions f such that $f(0) = 0$.

EXAMPLE 12. Let F be a field and let T be the linear operator on F^2 defined by

$$T(x_1, x_2) = (x_1 + x_2, x_1).$$

Then T is non-singular, because if $T(x_1, x_2) = 0$ we have

$$\begin{aligned} x_1 + x_2 &= 0 \\ x_1 &= 0 \end{aligned}$$

so that $x_1 = x_2 = 0$. We also see that T is onto; for, let (z_1, z_2) be any vector in F^2. To show that (z_1, z_2) is in the range of T we must find scalars x_1 and x_2 such that

$$\begin{aligned} x_1 + x_2 &= z_1 \\ x_1 &= z_2 \end{aligned}$$

and the obvious solution is $x_1 = z_2$, $x_2 = z_1 - z_2$. This last computation gives us an explicit formula for T^{-1}, namely,

$$T^{-1}(z_1, z_2) = (z_2, z_1 - z_2).$$

We have seen in Example 11 that a linear transformation may be non-singular without being onto and may be onto without being non-singular. The present example illustrates an important case in which that cannot happen.

Theorem 9. *Let* V *and* W *be finite-dimensional vector spaces over the field* F *such that dim* V = *dim* W. *If* T *is a linear transformation from* V *into* W, *the following are equivalent:*

(i) T *is invertible.*
(ii) T *is non-singular.*
(iii) T *is onto, that is, the range of* T *is* W.

Proof. Let $n = \dim V = \dim W$. From Theorem 2 we know that

$$\text{rank } (T) + \text{nullity } (T) = n.$$

Now T is non-singular if and only if nullity $(T) = 0$, and (since $n = \dim W$) the range of T is W if and only if rank $(T) = n$. Since the rank plus the nullity is n, the nullity is 0 precisely when the rank is n. Therefore T is non-singular if and only if $T(V) = W$. So, if either condition (ii) or (iii) holds, the other is satisfied as well and T is invertible. ∎

We caution the reader not to apply Theorem 9 except in the presence of finite-dimensionality and with $\dim V = \dim W$. Under the hypotheses of Theorem 9, the conditions (i), (ii), and (iii) are also equivalent to these.

(iv) *If* $\{\alpha_1, \ldots, \alpha_n\}$ *is basis for* V, *then* $\{T\alpha_1, \ldots, T\alpha_n\}$ *is a basis for* W.

(v) *There is some basis* $\{\alpha_1, \ldots, \alpha_n\}$ *for* V *such that* $\{T\alpha_1, \ldots, T\alpha_n\}$ *is a basis for* W.

We shall give a proof of the equivalence of the five conditions which contains a different proof that (i), (ii), and (iii) are equivalent.

(i) $\rightarrow$ (ii). If T is invertible, T is non-singular. (ii) $\rightarrow$ (iii). Suppose T is non-singular. Let $\{\alpha_1, \ldots, \alpha_n\}$ be a basis for V. By Theorem 8, $\{T\alpha_1, \ldots, T\alpha_n\}$ is a linearly independent set of vectors in W, and since the dimension of W is also n, this set of vectors is a basis for W. Now let β be any vector in W. There are scalars $c_1, \ldots, c_n$ such that

$$\begin{aligned}\beta &= c_1(T\alpha_1) + \cdots + c_n(T\alpha_n)\\ &= T(c_1\alpha_1 + \cdots + c_n\alpha_n)\end{aligned}$$

which shows that β is in the range of T. (iii) $\rightarrow$ (iv). We now assume that T is onto. If $\{\alpha_1, \ldots, \alpha_n\}$ is any basis for V, the vectors $T\alpha_1, \ldots, T\alpha_n$ span the range of T, which is all of W by assumption. Since the dimension of W is n, these n vectors must be linearly independent, that is, must comprise a basis for W. (iv) $\rightarrow$ (v). This requires no comment. (v) $\rightarrow$ (i). Suppose there is some basis $\{\alpha_1, \ldots, \alpha_n\}$ for V such that $\{T\alpha_1, \ldots, T\alpha_n\}$ is a basis for W. Since the $T\alpha_i$ span W, it is clear that the range of T is all of W. If $\alpha = c_1\alpha_1 + \cdots + c_n\alpha_n$ is in the null space of T, then

$$T(c_1\alpha_1 + \cdots + c_n\alpha_n) = 0$$

or

$$c_1(T\alpha_1) + \cdots + c_n(T\alpha_n) = 0$$

and since the $T\alpha_i$ are independent each $c_i = 0$, and thus $\alpha = 0$. We have shown that the range of T is W, and that T is non-singular, hence T is invertible.

The set of invertible linear operators on a space V, with the operation of composition, provides a nice example of what is known in algebra as a 'group.' Although we shall not have time to discuss groups in any detail, we shall at least give the definition.

Definition. *A* **group** *consists of the following.*

1. *A set* G;

2. *A rule (or operation) which associates with each pair of elements* x, y *in* G *an element* xy *in* G *in such a way that*

(a) x(yz) = (xy)z, *for all* x, y, *and* z *in* G (*associativity*);

(b) *there is an element* e *in* G *such that* ex = xe = x, *for every* x *in* G;

(c) *to each element* x *in* G *there corresponds an element* x^{-1} *in* G *such that* $xx^{-1} = x^{-1}x = e$.

We have seen that composition $(U, T) \rightarrow UT$ associates with each pair of invertible linear operators on a space V another invertible operator on V. Composition is an associative operation. The identity operator I

satisfies $IT = TI$ for each T, and for an invertible T there is (by Theorem 7) an invertible linear operator T^{-1} such that $TT^{-1} = T^{-1}T = I$. Thus the set of invertible linear operators on V, together with this operation, is a group. The set of invertible $n \times n$ matrices with matrix multiplication as the operation is another example of a group. A group is called **commutative** if it satisfies the condition $xy = yx$ for each x and y. The two examples we gave above are not commutative groups, in general. One often writes the operation in a commutative group as $(x, y) \to x + y$, rather than $(x, y) \to xy$, and then uses the symbol 0 for the 'identity' element e. The set of vectors in a vector space, together with the operation of vector addition, is a commutative group. A field can be described as a set with two operations, called addition and multiplication, which is a commutative group under addition, and in which the non-zero elements form a commutative group under multiplication, with the distributive law $x(y + z) = xy + xz$ holding.

Exercises

1. Let T and U be the linear operators on R^2 defined by

$$T(x_1, x_2) = (x_2, x_1) \quad \text{and} \quad U(x_1, x_2) = (x_1, 0).$$

(a) How would you describe T and U geometrically?

(b) Give rules like the ones defining T and U for each of the transformations $(U + T)$, UT, TU, T^2, U^2.

2. Let T be the (unique) linear operator on C^3 for which

$$T\epsilon_1 = (1, 0, i), \qquad T\epsilon_2 = (0, 1, 1), \qquad T\epsilon_3 = (i, 1, 0).$$

Is T invertible?

3. Let T be the linear operator on R^3 defined by

$$T(x_1, x_2, x_3) = (3x_1, x_1 - x_2, 2x_1 + x_2 + x_3).$$

Is T invertible? If so, find a rule for T^{-1} like the one which defines T.

4. For the linear operator T of Exercise 3, prove that

$$(T^2 - I)(T - 3I) = 0.$$

5. Let $C^{2\times 2}$ be the complex vector space of 2×2 matrices with complex entries. Let

$$B = \begin{bmatrix} 1 & -1 \\ -4 & 4 \end{bmatrix}$$

and let T be the linear operator on $C^{2\times 2}$ defined by $T(A) = BA$. What is the rank of T? Can you describe T^2?

6. Let T be a linear transformation from R^3 into R^2, and let U be a linear transformation from R^2 into R^3. Prove that the transformation UT is not invertible. Generalize the theorem.

7. Find two linear operators T and U on R^2 such that $TU = 0$ but $UT \neq 0$.

8. Let V be a vector space over the field F and T a linear operator on V. If $T^2 = 0$, what can you say about the relation of the range of T to the null space of T? Give an example of a linear operator T on R^2 such that $T^2 = 0$ but $T \neq 0$.

9. Let T be a linear operator on the finite-dimensional space V. Suppose there is a linear operator U on V such that $TU = I$. Prove that T is invertible and $U = T^{-1}$. Give an example which shows that this is false when V is not finite-dimensional. (*Hint:* Let $T = D$, the differentiation operator on the space of polynomial functions.)

10. Let A be an $m \times n$ matrix with entries in F and let T be the linear transformation from $F^{n\times 1}$ into $F^{m\times 1}$ defined by $T(X) = AX$. Show that if $m < n$ it may happen that T is onto without being non-singular. Similarly, show that if $m > n$ we may have T non-singular but not onto.

11. Let V be a finite-dimensional vector space and let T be a linear operator on V. Suppose that rank (T^2) = rank (T). Prove that the range and null space of T are disjoint, i.e., have only the zero vector in common.

12. Let p, m, and n be positive integers and F a field. Let V be the space of $m \times n$ matrices over F and W the space of $p \times n$ matrices over F. Let B be a fixed $p \times m$ matrix and let T be the linear transformation from V into W defined by $T(A) = BA$. Prove that T is invertible if and only if $p = m$ and B is an invertible $m \times m$ matrix.

3.3. Isomorphism

If V and W are vector spaces over the field F, any one-one linear transformation T of V onto W is called an **isomorphism of V onto W**. If there exists an isomorphism of V onto W, we say that V is **isomorphic** to W.

Note that V is trivially isomorphic to V, the identity operator being an isomorphism of V onto V. Also, if V is isomorphic to W via an isomorphism T, then W is isomorphic to V, because T^{-1} is an isomorphism of W onto V. The reader should find it easy to verify that if V is isomorphic to W and W is isomorphic to Z, then V is isomorphic to Z. Briefly, isomorphism is an equivalence relation on the class of vector spaces. If there exists an isomorphism of V onto W, we may sometimes say that V and W are isomorphic, rather than V is isomorphic to W. This will cause no confusion because V is isomorphic to W if and only if W is isomorphic to V.

Theorem 10. *Every* n-*dimensional vector space over the field* F *is isomorphic to the space* F^n.

Proof. Let V be an n-dimensional space over the field F and let $\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}$ be an ordered basis for V. We define a function T

from V into F^n, as follows: If α is in V, let $T\alpha$ be the n-tuple $(x_1, \ldots, x_n)$ of coordinates of α relative to the ordered basis $\mathcal{B}$, i.e., the n-tuple such that

$$\alpha = x_1\alpha_1 + \cdots + x_n\alpha_n.$$

In our discussion of coordinates in Chapter 2, we verified that this T is linear, one-one, and maps V onto F^n. ∎

For many purposes one often regards isomorphic vector spaces as being 'the same,' although the vectors and operations in the spaces may be quite different, that is, one often identifies isomorphic spaces. We shall not attempt a lengthy discussion of this idea at present but shall let the understanding of isomorphism and the sense in which isomorphic spaces are 'the same' grow as we continue our study of vector spaces.

We shall make a few brief comments. Suppose T is an isomorphism of V onto W. If S is a subset of V, then Theorem 8 tells us that S is linearly independent if and only if the set $T(S)$ in W is independent. Thus in deciding whether S is independent it doesn't matter whether we look at S or $T(S)$. From this one sees that an isomorphism is 'dimension preserving,' that is, any finite-dimensional subspace of V has the same dimension as its image under T. Here is a very simple illustration of this idea. Suppose A is an $m \times n$ matrix over the field F. We have really given two definitions of the solution space of the matrix A. The first is the set of all n-tuples $(x_1, \ldots, x_n)$ in F^n which satisfy each of the equations in the system $AX = 0$. The second is the set of all $n \times 1$ column matrices X such that $AX = 0$. The first solution space is thus a subspace of F^n and the second is a subspace of the space of all $n \times 1$ matrices over F. Now there is a completely obvious isomorphism between F^n and $F^{n\times 1}$, namely,

$$(x_1, \ldots, x_n) \to \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

Under this isomorphism, the first solution space of A is carried onto the second solution space. These spaces have the same dimension, and so if we want to prove a theorem about the dimension of the solution space, it is immaterial which space we choose to discuss. In fact, the reader would probably not balk if we chose to identify F^n and the space of $n \times 1$ matrices. We may do this when it is convenient, and when it is not convenient we shall not.

Exercises

1. Let V be the set of complex numbers and let F be the field of real numbers. With the usual operations, V is a vector space over F. Describe explicitly an isomorphism of this space onto R^2.

2. Let V be a vector space over the field of complex numbers, and suppose there is an isomorphism T of V onto C^3. Let $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ be vectors in V such that

$$T\alpha_1 = (1, 0, i), \qquad T\alpha_2 = (-2, 1 + i, 0),$$
$$T\alpha_3 = (-1, 1, 1), \qquad T\alpha_4 = (\sqrt{2}, i, 3).$$

(a) Is α_1 in the subspace spanned by α_2 and α_3?

(b) Let W_1 be the subspace spanned by α_1 and α_2, and let W_2 be the subspace spanned by α_3 and α_4. What is the intersection of W_1 and W_2?

(c) Find a basis for the subspace of V spanned by the four vectors α_j.

3. Let W be the set of all 2×2 complex Hermitian matrices, that is, the set of 2×2 complex matrices A such that $A_{ij} = \overline{A_{ji}}$ (the bar denoting complex conjugation). As we pointed out in Example 6 of Chapter 2, W is a vector space over the field of *real* numbers, under the usual operations. Verify that

$$(x, y, z, t) \to \begin{bmatrix} t + x & y + iz \\ y - iz & t - x \end{bmatrix}$$

is an isomorphism of R^4 onto W.

4. Show that $F^{m\times n}$ is isomorphic to F^{mn}.

5. Let V be the set of complex numbers regarded as a vector space over the field of real numbers (Exercise 1). We define a function T from V into the space of 2×2 real matrices, as follows. If $z = x + iy$ with x and y real numbers, then

$$T(z) = \begin{bmatrix} x + 7y & 5y \\ -10y & x - 7y \end{bmatrix}.$$

(a) Verify that T is a one-one (real) linear transformation of V into the space of 2×2 real matrices.

(b) Verify that $T(z_1z_2) = T(z_1)T(z_2)$.

(c) How would you describe the range of T?

6. Let V and W be finite-dimensional vector spaces over the field F. Prove that V and W are isomorphic if and only if $\dim V = \dim W$.

7. Let V and W be vector spaces over the field F and let U be an isomorphism of V onto W. Prove that $T \to UTU^{-1}$ is an isomorphism of $L(V, V)$ onto $L(W, W)$.

3.4. Representation of Transformations by Matrices

Let V be an n-dimensional vector space over the field F and let W be an m-dimensional vector space over F. Let $\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}$ be an ordered basis for V and $\mathcal{B}' = \{\beta_1, \ldots, \beta_m\}$ an ordered basis for W. If T is any linear transformation from V into W, then T is determined by its action on the vectors α_j. Each of the n vectors $T\alpha_j$ is uniquely expressible as a linear combination

$$T\alpha_j = \sum_{i=1}^{m} A_{ij}\beta_i \tag{3-3}$$

of the β_i, the scalars $A_{1j}, \ldots, A_{mj}$ being the coordinates of $T\alpha_j$ in the ordered basis $\mathcal{B}'$. Accordingly, the transformation T is determined by the mn scalars A_{ij} via the formulas (3-3). The $m \times n$ matrix A defined by $A(i, j) = A_{ij}$ is called **the matrix of T relative to the pair of ordered bases $\mathcal{B}$ and $\mathcal{B}'$**. Our immediate task is to understand explicitly how the matrix A determines the linear transformation T.

If $\alpha = x_1\alpha_1 + \cdots + x_n\alpha_n$ is a vector in V, then

$$\begin{aligned} T\alpha &= T\left(\sum_{j=1}^{n} x_j\alpha_j\right) \\ &= \sum_{j=1}^{n} x_j(T\alpha_j) \\ &= \sum_{j=1}^{n} x_j \sum_{i=1}^{m} A_{ij}\beta_i \\ &= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij}x_j\right) \beta_i. \end{aligned}$$

If X is the coordinate matrix of α in the ordered basis $\mathcal{B}$, then the computation above shows that AX is the coordinate matrix of the vector $T\alpha$ in the ordered basis $\mathcal{B}'$, because the scalar

$$\sum_{j=1}^{n} A_{ij}x_j$$

is the entry in the ith row of the column matrix AX. Let us also observe that if A is any $m \times n$ matrix over the field F, then

$$T\left(\sum_{j=1}^{n} x_j\alpha_j\right) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij}x_j\right) \beta_i \tag{3-4}$$

defines a linear transformation T from V into W, the matrix of which is A, relative to $\mathcal{B}$, $\mathcal{B}'$. We summarize formally:

Theorem 11. *Let* V *be an* n*-dimensional vector space over the field* F *and* W *an* m*-dimensional vector space over* F. *Let* $\mathcal{B}$ *be an ordered basis for* V *and* $\mathcal{B}'$ *an ordered basis for* W. *For each linear transformation* T *from* V *into* W, *there is an* m $\times$ n *matrix* A *with entries in* F *such that*

$$[\mathrm{T}\alpha]_{\mathcal{B}'} = \mathrm{A}[\alpha]_{\mathcal{B}}$$

for every vector α *in* V. *Furthermore,* $\mathrm{T} \rightarrow \mathrm{A}$ *is a one-one correspondence between the set of all linear transformations from* V *into* W *and the set of all* m $\times$ n *matrices over the field* F.

The matrix A which is associated with T in Theorem 11 is called the **matrix of T relative to the ordered bases $\mathcal{B}$, $\mathcal{B}'$**. Note that Equation (3-3) says that A is the matrix whose columns $A_1, \ldots, A_n$ are given by

$$A_j = [T\alpha_j]_{\mathcal{B}'}, \qquad j = 1, \ldots, n.$$

If U is another linear transformation from V into W and $B = [B_1, \ldots, B_n]$ is the matrix of U relative to the ordered bases $\mathcal{B}$, $\mathcal{B}'$ then $cA + B$ is the matrix of $cT + U$ relative to $\mathcal{B}$, $\mathcal{B}'$. That is clear because

$$\begin{aligned} cA_j + B_j &= c[T\alpha_j]_{\mathcal{B}'} + [U\alpha_j]_{\mathcal{B}'} \\ &= [cT\alpha_j + U\alpha_j]_{\mathcal{B}'} \\ &= [(cT + U)\alpha_j]_{\mathcal{B}'}. \end{aligned}$$

Theorem 12. *Let* V *be an* n*-dimensional vector space over the field* F *and let* W *be an* m*-dimensional vector space over* F. *For each pair of ordered bases* $\mathcal{B}$, $\mathcal{B}'$ *for* V *and* W *respectively, the function which assigns to a linear transformation* T *its matrix relative to* $\mathcal{B}$, $\mathcal{B}'$ *is an isomorphism between the space* L(V, W) *and the space of all* m $\times$ n *matrices over the field* F.

Proof. We observed above that the function in question is linear, and as stated in Theorem 11, this function is one-one and maps $L(V, W)$ onto the set of $m \times n$ matrices. ∎

We shall be particularly interested in the representation by matrices of linear transformations of a space into itself, i.e., linear operators on a space V. In this case it is most convenient to use the same ordered basis in each case, that is, to take $\mathcal{B} = \mathcal{B}'$. We shall then call the representing matrix simply the **matrix of T relative to the ordered basis** $\mathcal{B}$. Since this concept will be so important to us, we shall review its definition. If T is a linear operator on the finite-dimensional vector space V and $\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}$ is an ordered basis for V, the matrix of T relative to $\mathcal{B}$ (or, the matrix of T in the ordered basis $\mathcal{B}$) is the $n \times n$ matrix A whose entries A_{ij} are defined by the equations

$$T\alpha_j = \sum_{i=1}^{n} A_{ij}\alpha_i, \qquad j = 1, \ldots, n. \tag{3-5}$$

One must always remember that this matrix representing T depends upon the ordered basis $\mathcal{B}$, and that there is a representing matrix for T in each ordered basis for V. (For transformations of one space into another the matrix depends upon two ordered bases, one for V and one for W.) In order that we shall not forget this dependence, we shall use the notation

$$[T]_{\mathcal{B}}$$

for the matrix of the linear operator T in the ordered basis $\mathcal{B}$. The manner in which this matrix and the ordered basis describe T is that for each α in V

$$[T\alpha]_{\mathcal{B}} = [T]_{\mathcal{B}}[\alpha]_{\mathcal{B}}.$$

EXAMPLE 13. Let V be the space of $n \times 1$ column matrices over the field F; let W be the space of $m \times 1$ matrices over F; and let A be a fixed $m \times n$ matrix over F. Let T be the linear transformation of V into W defined by $T(X) = AX$. Let $\mathcal{B}$ be the ordered basis for V analogous to the

standard basis in F^n, i.e., the ith vector in $\mathcal{B}$ in the $n \times 1$ matrix X_i with a 1 in row i and all other entries 0. Let $\mathcal{B}'$ be the corresponding ordered basis for W, i.e., the jth vector in $\mathcal{B}'$ is the $m \times 1$ matrix Y_j with a 1 in row j and all other entries 0. Then the matrix of T relative to the pair $\mathcal{B}$, $\mathcal{B}'$ is the matrix A itself. This is clear because the matrix AX_j is the jth column of A.

EXAMPLE 14. Let F be a field and let T be the operator on F^2 defined by

$$T(x_1, x_2) = (x_1, 0).$$

It is easy to see that T is a linear operator on F^2. Let $\mathcal{B}$ be the standard ordered basis for F^2, $\mathcal{B} = \{\epsilon_1, \epsilon_2\}$. Now

$$\begin{aligned} T\epsilon_1 &= T(1, 0) = (1, 0) = 1\epsilon_1 + 0\epsilon_2 \\ T\epsilon_2 &= T(0, 1) = (0, 0) = 0\epsilon_1 + 0\epsilon_2 \end{aligned}$$

so the matrix of T in the ordered basis $\mathcal{B}$ is

$$[T]_{\mathcal{B}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

EXAMPLE 15. Let V be the space of all polynomial functions from R into R of the form

$$f(x) = c_0 + c_1x + c_2x^2 + c_3x^3$$

that is, the space of polynomial functions of degree three or less. The differentiation operator D of Example 2 maps V into V, since D is 'degree decreasing.' Let $\mathcal{B}$ be the ordered basis for V consisting of the four functions f_1, f_2, f_3, f_4 defined by $f_j(x) = x^{j-1}$. Then

$$\begin{aligned} (Df_1)(x) &= 0, & Df_1 &= 0f_1 + 0f_2 + 0f_3 + 0f_4 \\ (Df_2)(x) &= 1, & Df_2 &= 1f_1 + 0f_2 + 0f_3 + 0f_4 \\ (Df_3)(x) &= 2x, & Df_3 &= 0f_1 + 2f_2 + 0f_3 + 0f_4 \\ (Df_4)(x) &= 3x^2, & Df_4 &= 0f_1 + 0f_2 + 3f_3 + 0f_4 \end{aligned}$$

so that the matrix of D in the ordered basis $\mathcal{B}$ is

$$[D]_{\mathcal{B}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

We have seen what happens to representing matrices when transformations are added, namely, that the matrices add. We should now like to ask what happens when we compose transformations. More specifically, let V, W, and Z be vector spaces over the field F of respective dimensions n, m, and p. Let T be a linear transformation from V into W and U a linear transformation from W into Z. Suppose we have ordered bases

$$\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}, \qquad \mathcal{B}' = \{\beta_1, \ldots, \beta_m\}, \qquad \mathcal{B}'' = \{\gamma_1, \ldots, \gamma_p\}$$

for the respective spaces V, W, and Z. Let A be the matrix of T relative to the pair $\mathcal{B}$, $\mathcal{B}'$ and let B be the matrix of U relative to the pair $\mathcal{B}'$, $\mathcal{B}''$. It is then easy to see that the matrix C of the transformation UT relative to the pair $\mathcal{B}$, $\mathcal{B}''$ is the product of B and A; for, if α is any vector in V

$$\begin{aligned}[T\alpha]_{\mathcal{B}'} &= A[\alpha]_{\mathcal{B}} \\ [U(T\alpha)]_{\mathcal{B}''} &= B[T\alpha]_{\mathcal{B}'}\end{aligned}$$

and so

$$[(UT)(\alpha)]_{\mathcal{B}''} = BA[\alpha]_{\mathcal{B}}$$

and hence, by the definition and uniqueness of the representing matrix, we must have $C = BA$. One can also see this by carrying out the computation

$$\begin{aligned}(UT)(\alpha_j) &= U(T\alpha_j) \\ &= U\left(\sum_{k=1}^{m} A_{kj}\beta_k\right) \\ &= \sum_{k=1}^{m} A_{kj}(U\beta_k) \\ &= \sum_{k=1}^{m} A_{kj} \sum_{i=1}^{p} B_{ik}\gamma_i \\ &= \sum_{i=1}^{p} \left(\sum_{k=1}^{m} B_{ik}A_{kj}\right)\gamma_i\end{aligned}$$

so that we must have

$$C_{ij} = \sum_{k=1}^{m} B_{ik}A_{kj}. \tag{3-6}$$

We motivated the definition (3-6) of matrix multiplication via operations on the rows of a matrix. One sees here that a very strong motivation for the definition is to be found in composing linear transformations. Let us summarize formally.

Theorem 13. *Let* V, W, *and* Z *be finite-dimensional vector spaces over the field* F; *let* T *be a linear transformation from* V *into* W *and* U *a linear transformation from* W *into* Z. *If* $\mathcal{B}$, $\mathcal{B}'$, *and* $\mathcal{B}''$ *are ordered bases for the spaces* V, W, *and* Z, *respectively, if* A *is the matrix of* T *relative to the pair* $\mathcal{B}$, $\mathcal{B}'$, *and* B *is the matrix of* U *relative to the pair* $\mathcal{B}'$, $\mathcal{B}''$, *then the matrix of the composition* UT *relative to the pair* $\mathcal{B}$, $\mathcal{B}''$ *is the product matrix* C = BA.

We remark that Theorem 13 gives a proof that matrix multiplication is associative—a proof which requires no calculations and is independent of the proof we gave in Chapter 1. We should also point out that we proved a special case of Theorem 13 in Example 12.

It is important to note that if T and U are linear operators on a space V and we are representing by a single ordered basis $\mathcal{B}$, then Theorem 13 assumes the simple form $[UT]_{\mathcal{B}} = [U]_{\mathcal{B}}[T]_{\mathcal{B}}$. Thus in this case, the

correspondence which $\mathcal{B}$ determines between operators and matrices is not only a vector space isomorphism but also preserves products. A simple consequence of this is that the linear operator T is invertible if and only if $[T]_{\mathcal{B}}$ is an invertible matrix. For, the identity operator I is represented by the identity matrix in any ordered basis, and thus

$$UT = TU = I$$

is equivalent to

$$[U]_{\mathcal{B}}[T]_{\mathcal{B}} = [T]_{\mathcal{B}}[U]_{\mathcal{B}} = I.$$

Of course, when T is invertible

$$[T^{-1}]_{\mathcal{B}} = [T]_{\mathcal{B}}^{-1}.$$

Now we should like to inquire what happens to representing matrices when the ordered basis is changed. For the sake of simplicity, we shall consider this question only for linear operators on a space V, so that we can use a single ordered basis. The specific question is this. Let T be a linear operator on the finite-dimensional space V, and let

$$\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\} \quad \text{and} \quad \mathcal{B}' = \{\alpha_1', \ldots, \alpha_n'\}$$

be two ordered bases for V. How are the matrices $[T]_{\mathcal{B}}$ and $[T]_{\mathcal{B}'}$ related? As we observed in Chapter 2, there is a unique (invertible) $n \times n$ matrix P such that

$$[\alpha]_{\mathcal{B}} = P[\alpha]_{\mathcal{B}'} \tag{3-7}$$

for every vector α in V. It is the matrix $P = [P_1, \ldots, P_n]$ where $P_j = [\alpha_j']_{\mathcal{B}}$. By definition

$$[T\alpha]_{\mathcal{B}} = [T]_{\mathcal{B}}[\alpha]_{\mathcal{B}}. \tag{3-8}$$

Applying (3-7) to the vector $T\alpha$, we have

$$[T\alpha]_{\mathcal{B}} = P[T\alpha]_{\mathcal{B}'}. \tag{3-9}$$

Combining (3-7), (3-8), and (3-9), we obtain

$$[T]_{\mathcal{B}}P[\alpha]_{\mathcal{B}'} = P[T\alpha]_{\mathcal{B}'}$$

or

$$P^{-1}[T]_{\mathcal{B}}P[\alpha]_{\mathcal{B}'} = [T\alpha]_{\mathcal{B}'}$$

and so it must be that

$$[T]_{\mathcal{B}'} = P^{-1}[T]_{\mathcal{B}}P. \tag{3-10}$$

This answers our question.

Before stating this result formally, let us observe the following. There is a unique linear operator U which carries $\mathcal{B}$ onto $\mathcal{B}'$, defined by

$$U\alpha_j = \alpha_j', \qquad j = 1, \ldots, n.$$

This operator U is invertible since it carries a basis for V onto a basis for

V. The matrix P (above) is precisely the matrix of the operator U in the ordered basis $\mathcal{B}$. For, P is defined by

$$\alpha_j' = \sum_{i=1}^{n} P_{ij}\alpha_i$$

and since $U\alpha_j = \alpha_j'$, this equation can be written

$$U\alpha_j = \sum_{i=1}^{n} P_{ij}\alpha_i.$$

So $P = [U]_{\mathcal{B}}$, by definition.

Theorem 14. *Let* V *be a finite-dimensional vector space over the field* F, *and let*

$$\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\} \quad \textit{and} \quad \mathcal{B}' = \{\alpha_1', \ldots, \alpha_n'\}$$

be ordered bases for V. *Suppose* T *is a linear operator on* V. *If* $P = [P_1, \ldots, P_n]$ *is the* $n \times n$ *matrix with columns* $P_j = [\alpha_j']_{\mathcal{B}}$, *then*

$$[T]_{\mathcal{B}'} = P^{-1}[T]_{\mathcal{B}}P.$$

Alternatively, if U *is the invertible operator on* V *defined by* $U\alpha_j = \alpha_j'$, $j = 1, \ldots, n$, *then*

$$[T]_{\mathcal{B}'} = [U]_{\mathcal{B}}^{-1}[T]_{\mathcal{B}}[U]_{\mathcal{B}}.$$

EXAMPLE 16. Let T be the linear operator on R^2 defined by $T(x_1, x_2) = (x_1, 0)$. In Example 14 we showed that the matrix of T in the standard ordered basis $\mathcal{B} = \{\epsilon_1, \epsilon_2\}$ is

$$[T]_{\mathcal{B}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Suppose $\mathcal{B}'$ is the ordered basis for R^2 consisting of the vectors $\epsilon_1' = (1, 1)$, $\epsilon_2' = (2, 1)$. Then

$$\begin{aligned} \epsilon_1' &= \epsilon_1 + \epsilon_2 \\ \epsilon_2' &= 2\epsilon_1 + \epsilon_2 \end{aligned}$$

so that P is the matrix

$$P = \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix}.$$

By a short computation

$$P^{-1} = \begin{bmatrix} -1 & 2 \\ 1 & -1 \end{bmatrix}.$$

Thus

$$\begin{aligned} [T]_{\mathcal{B}'} &= P^{-1}[T]_{\mathcal{B}}P \\ &= \begin{bmatrix} -1 & 2 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} -1 & 2 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} -1 & -2 \\ 1 & 2 \end{bmatrix}. \end{aligned}$$

We can easily check that this is correct because

$$\begin{aligned} T\epsilon_1' &= (1, 0) = -\epsilon_1' + \epsilon_2' \\ T\epsilon_2' &= (2, 0) = -2\epsilon_1' + 2\epsilon_2'. \end{aligned}$$

EXAMPLE 17. Let V be the space of polynomial functions from R into R which have 'degree' less than or equal to 3. As in Example 15, let D be the differentiation operator on V, and let

$$\mathcal{B} = \{f_1, f_2, f_3, f_4\}$$

be the ordered basis for V defined by $f_i(x) = x^{i-1}$. Let t be a real number and define $g_i(x) = (x + t)^{i-1}$, that is

$$\begin{aligned} g_1 &= f_1 \\ g_2 &= tf_1 + f_2 \\ g_3 &= t^2f_1 + 2tf_2 + f_3 \\ g_4 &= t^3f_1 + 3t^2f_2 + 3tf_3 + f_4. \end{aligned}$$

Since the matrix

$$P = \begin{bmatrix} 1 & t & t^2 & t^3 \\ 0 & 1 & 2t & 3t^2 \\ 0 & 0 & 1 & 3t \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

is easily seen to be invertible with

$$P^{-1} = \begin{bmatrix} 1 & -t & t^2 & -t^3 \\ 0 & 1 & -2t & 3t^2 \\ 0 & 0 & 1 & -3t \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

it follows that $\mathcal{B}' = \{g_1, g_2, g_3, g_4\}$ is an ordered basis for V. In Example 15, we found that the matrix of D in the ordered basis $\mathcal{B}$ is

$$[D]_{\mathcal{B}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The matrix of D in the ordered basis $\mathcal{B}'$ is thus

$$\begin{aligned} P^{-1}[D]_{\mathcal{B}}P &= \begin{bmatrix} 1 & -t & t^2 & t^3 \\ 0 & 1 & -2t & 3t^2 \\ 0 & 0 & 1 & -3t \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & t & t^2 & t^3 \\ 0 & 1 & 2t & 3t^2 \\ 0 & 0 & 1 & 3t \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -t & t^2 & t^3 \\ 0 & 1 & -2t & 3t^2 \\ 0 & 0 & 1 & -3t \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 2t & 3t^2 \\ 0 & 0 & 2 & 6t \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Thus D is represented by the same matrix in the ordered bases $\mathcal{B}$ and $\mathcal{B}'$. Of course, one can see this somewhat more directly since

$$\begin{aligned} Dg_1 &= 0 \\ Dg_2 &= g_1 \\ Dg_3 &= 2g_2 \\ Dg_4 &= 3g_3. \end{aligned}$$

This example illustrates a good point. If one knows the matrix of a linear operator in some ordered basis $\mathcal{B}$ and wishes to find the matrix in another ordered basis $\mathcal{B}'$, it is often most convenient to perform the coordinate change using the invertible matrix P; however, it may be a much simpler task to find the representing matrix by a direct appeal to its definition.

Definition. *Let* A *and* B *be* n $\times$ n *(square) matrices over the field* F. *We say that* B **is similar to** A **over** F *if there is an invertible* n $\times$ n *matrix* P *over* F *such that* B = $P^{-1}AP$.

According to Theorem 14, we have the following: If V is an n-dimensional vector space over F and $\mathcal{B}$ and $\mathcal{B}'$ are two ordered bases for V, then for each linear operator T on V the matrix $B = [T]_{\mathcal{B}'}$ is similar to the matrix $A = [T]_{\mathcal{B}}$. The argument also goes in the other direction. Suppose A and B are $n \times n$ matrices and that B is similar to A. Let V be any n-dimensional space over F and let $\mathcal{B}$ be an ordered basis for V. Let T be the linear operator on V which is represented in the basis $\mathcal{B}$ by A. If $B = P^{-1}AP$, let $\mathcal{B}'$ be the ordered basis for V obtained from $\mathcal{B}$ by P, i.e.,

$$\alpha_j' = \sum_{i=1}^{n} P_{ij}\alpha_i.$$

Then the matrix of T in the ordered basis $\mathcal{B}'$ will be B.

Thus the statement that B is similar to A means that on each n-dimensional space over F the matrices A and B represent the same linear transformation in two (possibly) different ordered bases.

Note that each $n \times n$ matrix A is similar to itself, using $P = I$; if B is similar to A, then A is similar to B, for $B = P^{-1}AP$ implies that $A = (P^{-1})^{-1}BP^{-1}$; if B is similar to A and C is similar to B, then C is similar to A, for $B = P^{-1}AP$ and $C = Q^{-1}BQ$ imply that $C = (PQ)^{-1}A(PQ)$. Thus, similarity is an equivalence relation on the set of $n \times n$ matrices over the field F. Also note that the only matrix similar to the identity matrix I is I itself, and that the only matrix similar to the zero matrix is the zero matrix itself.

Exercises

1. Let T be the linear operator on C^2 defined by $T(x_1, x_2) = (x_1, 0)$. Let $\mathcal{B}$ be the standard ordered basis for C^2 and let $\mathcal{B}' = \{\alpha_1, \alpha_2\}$ be the ordered basis defined by $\alpha_1 = (1, i)$, $\alpha_2 = (-i, 2)$.

(a) What is the matrix of T relative to the pair $\mathcal{B}$, $\mathcal{B}'$?

(b) What is the matrix of T relative to the pair $\mathcal{B}'$, $\mathcal{B}$?

(c) What is the matrix of T in the ordered basis $\mathcal{B}'$?

(d) What is the matrix of T in the ordered basis $\{\alpha_2, \alpha_1\}$?

2. Let T be the linear transformation from R^3 into R^2 defined by

$$T(x_1, x_2, x_3) = (x_1 + x_2, 2x_3 - x_1).$$

(a) If $\mathcal{B}$ is the standard ordered basis for R^3 and $\mathcal{B}'$ is the standard ordered basis for R^2, what is the matrix of T relative to the pair $\mathcal{B}$, $\mathcal{B}'$?

(b) If $\mathcal{B} = \{\alpha_1, \alpha_2, \alpha_3\}$ and $\mathcal{B}' = \{\beta_1, \beta_2\}$, where

$$\alpha_1 = (1, 0, -1), \quad \alpha_2 = (1, 1, 1), \quad \alpha_3 = (1, 0, 0), \quad \beta_1 = (0, 1), \quad \beta_2 = (1, 0)$$

what is the matrix of T relative to the pair $\mathcal{B}$, $\mathcal{B}'$?

3. Let T be a linear operator on F^n, let A be the matrix of T in the standard ordered basis for F^n, and let W be the subspace of F^n spanned by the column vectors of A. What does W have to do with T?

4. Let V be a two-dimensional vector space over the field F, and let $\mathcal{B}$ be an ordered basis for V. If T is a linear operator on V and

$$[T]_{\mathcal{B}} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

prove that $T^2 - (a + d)T + (ad - bc)I = 0$.

5. Let T be the linear operator on R^3, the matrix of which in the standard ordered basis is

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ -1 & 3 & 4 \end{bmatrix}.$$

Find a basis for the range of T and a basis for the null space of T.

6. Let T be the linear operator on R^2 defined by

$$T(x_1, x_2) = (-x_2, x_1).$$

(a) What is the matrix of T in the standard ordered basis for R^2?

(b) What is the matrix of T in the ordered basis $\mathcal{B} = \{\alpha_1, \alpha_2\}$, where $\alpha_1 = (1, 2)$ and $\alpha_2 = (1, -1)$?

(c) Prove that for every real number c the operator $(T - cI)$ is invertible.

(d) Prove that if $\mathcal{B}$ is any ordered basis for R^2 and $[T]_{\mathcal{B}} = A$, then $A_{12}A_{21} \neq 0$.

7. Let T be the linear operator on R^3 defined by

$$T(x_1, x_2, x_3) = (3x_1 + x_3, -2x_1 + x_2, -x_1 + 2x_2 + 4x_3).$$

(a) What is the matrix of T in the standard ordered basis for R^3?

(b) What is the matrix of T in the ordered basis

$$\{\alpha_1, \alpha_2, \alpha_3\}$$

where $\alpha_1 = (1, 0, 1)$, $\alpha_2 = (-1, 2, 1)$, and $\alpha_3 = (2, 1, 1)$?

(c) Prove that T is invertible and give a rule for T^{-1} like the one which defines T.

8. Let θ be a real number. Prove that the following two matrices are similar over the field of complex numbers:

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad \begin{bmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{bmatrix}$$

(*Hint:* Let T be the linear operator on C^2 which is represented by the first matrix in the standard ordered basis. Then find vectors α_1 and α_2 such that $T\alpha_1 = e^{i\theta}\alpha_1$, $T\alpha_2 = e^{-i\theta}\alpha_2$, and $\{\alpha_1, \alpha_2\}$ is a basis.)

9. Let V be a finite-dimensional vector space over the field F and let S and T be linear operators on V. We ask: When do there exist ordered bases $\mathcal{B}$ and $\mathcal{B}'$ for V such that $[S]_{\mathcal{B}} = [T]_{\mathcal{B}'}$? Prove that such bases exist if and only if there is an invertible linear operator U on V such that $T = USU^{-1}$. (*Outline of proof:* If $[S]_{\mathcal{B}} = [T]_{\mathcal{B}'}$, let U be the operator which carries $\mathcal{B}$ onto $\mathcal{B}'$ and show that $S = UTU^{-1}$. Conversely, if $T = USU^{-1}$ for some invertible U, let $\mathcal{B}$ be any ordered basis for V and let $\mathcal{B}'$ be its image under U. Then show that $[S]_{\mathcal{B}} = [T]_{\mathcal{B}'}$.)

10. We have seen that the linear operator T on R^2 defined by $T(x_1, x_2) = (x_1, 0)$ is represented in the standard ordered basis by the matrix

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

This operator satisfies $T^2 = T$. Prove that if S is a linear operator on R^2 such that $S^2 = S$, then $S = 0$, or $S = I$, or there is an ordered basis $\mathcal{B}$ for R^2 such that $[S]_{\mathcal{B}} = A$ (above).

11. Let W be the space of all $n \times 1$ column matrices over a field F. If A is an $n \times n$ matrix over F, then A defines a linear operator L_A on W through left multiplication: $L_A(X) = AX$. Prove that every linear operator on W is left multiplication by some $n \times n$ matrix, i.e., is L_A for some A.

Now suppose V is an n-dimensional vector space over the field F, and let $\mathcal{B}$ be an ordered basis for V. For each α in V, define $U\alpha = [\alpha]_{\mathcal{B}}$. Prove that U is an isomorphism of V onto W. If T is a linear operator on V, then UTU^{-1} is a linear operator on W. Accordingly, UTU^{-1} is left multiplication by some $n \times n$ matrix A. What is A?

12. Let V be an n-dimensional vector space over the field F, and let $\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}$ be an ordered basis for V.

(a) According to Theorem 1, there is a unique linear operator T on V such that

$$T\alpha_j = \alpha_{j+1}, \qquad j = 1, \ldots, n-1, \qquad T\alpha_n = 0.$$

What is the matrix A of T in the ordered basis $\mathcal{B}$?

(b) Prove that $T^n = 0$ but $T^{n-1} \neq 0$.

(c) Let S be any linear operator on V such that $S^n = 0$ but $S^{n-1} \neq 0$. Prove that there is an ordered basis $\mathcal{B}'$ for V such that the matrix of S in the ordered basis $\mathcal{B}'$ is the matrix A of part (a).

(d) Prove that if M and N are $n \times n$ matrices over F such that $M^n = N^n = 0$ but $M^{n-1} \neq 0 \neq N^{n-1}$, then M and N are similar.

13. Let V and W be finite-dimensional vector spaces over the field F and let T be a linear transformation from V into W. If

$$\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\} \quad \text{and} \quad \mathcal{B}' = \{\beta_1, \ldots, \beta_m\}$$

are ordered bases for V and W, respectively, define the linear transformations $E^{p,q}$ as in the proof of Theorem 5: $E^{p,q}(\alpha_i) = \delta_{iq}\beta_p$. Then the $E^{p,q}$, $1 \leq p \leq m$, $1 \leq q \leq n$, form a basis for $L(V, W)$, and so

$$T = \sum_{p=1}^{m} \sum_{q=1}^{n} A_{pq}E^{p,q}$$

for certain scalars A_{pq} (the coordinates of T in this basis for $L(V, W)$). Show that the matrix A with entries $A(p, q) = A_{pq}$ is precisely the matrix of T relative to the pair $\mathcal{B}$, $\mathcal{B}'$.

3.5. Linear Functionals

If V is a vector space over the field F, a linear transformation f from V into the scalar field F is also called a **linear functional** on V. If we start from scratch, this means that f is a function from V into F such that

$$f(c\alpha + \beta) = cf(\alpha) + f(\beta)$$

for all vectors α and β in V and all scalars c in F. The concept of linear functional is important in the study of finite-dimensional spaces because it helps to organize and clarify the discussion of subspaces, linear equations, and coordinates.

EXAMPLE 18. Let F be a field and let $a_1, \ldots, a_n$ be scalars in F. Define a function f on F^n by

$$f(x_1, \ldots, x_n) = a_1x_1 + \cdots + a_nx_n.$$

Then f is a linear functional on F^n. It is the linear functional which is represented by the matrix $[a_1 \cdots a_n]$ relative to the standard ordered basis for F^n and the basis $\{1\}$ for F:

$$a_j = f(\epsilon_j), \qquad j = 1, \ldots, n.$$

Every linear functional on F^n is of this form, for some scalars $a_1, \ldots, a_n$. That is immediate from the definition of linear functional because we define $a_j = f(\epsilon_j)$ and use the linearity

$$\begin{aligned} f(x_1, \ldots, x_n) &= f\left(\sum_j x_j\epsilon_j\right) \\ &= \sum_j x_jf(\epsilon_j) \\ &= \sum_j a_jx_j. \end{aligned}$$

EXAMPLE 19. Here is an important example of a linear functional. Let n be a positive integer and F a field. If A is an $n \times n$ matrix with entries in F, the **trace** of A is the scalar

$$\operatorname{tr} A = A_{11} + A_{22} + \cdots + A_{nn}.$$

The trace function is a linear functional on the matrix space $F^{n\times n}$ because

$$\begin{aligned}\operatorname{tr}(cA + B) &= \sum_{i=1}^{n} (cA_{ii} + B_{ii}) \\ &= c \sum_{i=1}^{n} A_{ii} + \sum_{i=1}^{n} B_{ii} \\ &= c \operatorname{tr} A + \operatorname{tr} B.\end{aligned}$$

EXAMPLE 20. Let V be the space of all polynomial functions from the field F into itself. Let t be an element of F. If we define

$$L_t(p) = p(t)$$

then L_t is a linear functional on V. One usually describes this by saying that, for each t, 'evaluation at t' is a linear functional on the space of polynomial functions. Perhaps we should remark that the fact that the functions are polynomials plays no role in this example. Evaluation at t is a linear functional on the space of all functions from F into F.

EXAMPLE 21. This may be the most important linear functional in mathematics. Let $[a, b]$ be a closed interval on the real line and let $C([a, b])$ be the space of continuous real-valued functions on $[a, b]$. Then

$$L(g) = \int_a^b g(t)\, dt$$

defines a linear functional L on $C([a, b])$.

If V is a vector space, the collection of all linear functionals on V forms a vector space in a natural way. It is the space $L(V, F)$. We denote this space by V^* and call it the **dual space** of V:

$$V^* = L(V, F).$$

If V is finite-dimensional, we can obtain a rather explicit description of the dual space V^*. From Theorem 5 we know something about the space V^*, namely that

$$\dim V^* = \dim V.$$

Let $\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}$ be a basis for V. According to Theorem 1, there is (for each i) a unique linear functional f_i on V such that

$$f_i(\alpha_j) = \delta_{ij}. \tag{3-11}$$

In this way we obtain from $\mathcal{B}$ a set of n distinct linear functionals $f_1, \ldots, f_n$ on V. These functionals are also linearly independent. For, suppose

(3-12) $$f = \sum_{i=1}^{n} c_i f_i.$$

Then

$$\begin{aligned} f(\alpha_j) &= \sum_{i=1}^{n} c_i f_i(\alpha_j) \\ &= \sum_{i=1}^{n} c_i \delta_{ij} \\ &= c_j. \end{aligned}$$

In particular, if f is the zero functional, $f(\alpha_j) = 0$ for each j and hence the scalars c_j are all 0. Now $f_1, \ldots, f_n$ are n linearly independent functionals, and since we know that V^* has dimension n, it must be that $\mathcal{B}^* = \{f_1, \ldots, f_n\}$ is a basis for V^*. This basis is called the **dual basis** of $\mathcal{B}$.

Theorem 15. *Let* V *be a finite-dimensional vector space over the field* F, *and let* $\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}$ *be a basis for* V. *Then there is a unique dual basis* $\mathcal{B}^* = \{f_1, \ldots, f_n\}$ *for* V^* *such that* $f_i(\alpha_j) = \delta_{ij}$. *For each linear functional* f *on* V *we have*

(3-13) $$f = \sum_{i=1}^{n} f(\alpha_i) f_i$$

and for each vector α *in* V *we have*

(3-14) $$\alpha = \sum_{i=1}^{n} f_i(\alpha)\alpha_i.$$

Proof. We have shown above that there is a unique basis which is 'dual' to $\mathcal{B}$. If f is a linear functional on V, then f is some linear combination (3-12) of the f_i, and as we observed after (3-12) the scalars c_j must be given by $c_j = f(\alpha_j)$. Similarly, if

$$\alpha = \sum_{i=1}^{n} x_i \alpha_i$$

is a vector in V, then

$$\begin{aligned} f_j(\alpha) &= \sum_{i=1}^{n} x_i f_j(\alpha_i) \\ &= \sum_{i=1}^{n} x_i \delta_{ij} \\ &= x_j \end{aligned}$$

so that the unique expression for α as a linear combination of the α_i is

$$\alpha = \sum_{i=1}^{n} f_i(\alpha)\alpha_i. \quad \blacksquare$$

Equation (3-14) provides us with a nice way of describing what the dual basis is. It says, if $\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}$ is an ordered basis for V and

$\mathcal{B}^* = \{f_1, \ldots, f_n\}$ is the dual basis, then f_i is precisely the function which assigns to each vector α in V the ith coordinate of α relative to the ordered basis $\mathcal{B}$. Thus we may also call the f_i the coordinate functions for $\mathcal{B}$. The formula (3-13), when combined with (3-14) tells us the following: If f is in V^*, and we let $f(\alpha_i) = \alpha_i$, then when

$$\alpha = x_1\alpha_1 + \cdots + x_n\alpha_n$$

we have

$$f(\alpha) = a_1x_1 + \cdots + a_nx_n. \tag{3-15}$$

In other words, if we choose an ordered basis $\mathcal{B}$ for V and describe each vector in V by its n-tuple of coordinates $(x_1, \ldots, x_n)$ relative to $\mathcal{B}$, then every linear functional on V has the form (3-15). This is the natural generalization of Example 18, which is the special case $V = F^n$ and $\mathcal{B} = \{\epsilon_1, \ldots, \epsilon_n\}$.

EXAMPLE 22. Let V be the vector space of all polynomial functions from R into R which have degree less than or equal to 2. Let t_1, t_2, and t_3 be any three *distinct* real numbers, and let

$$L_i(p) = p(t_i).$$

Then L_1, L_2, and L_3 are linear functionals on V. These functionals are linearly independent; for, suppose

$$L = c_1L_1 + c_2L_2 + c_3L_3.$$

If $L = 0$, i.e., if $L(p) = 0$ for each p in V, then applying L to the particular polynomial 'functions' 1, x, x^2, we obtain

$$\begin{aligned} c_1 + c_2 + c_3 &= 0 \\ t_1c_1 + t_2c_2 + t_3c_3 &= 0 \\ t_1^2c_1 + t_2^2c_2 + t_3^2c_3 &= 0 \end{aligned}$$

From this it follows that $c_1 = c_2 = c_3 = 0$, because (as a short computation shows) the matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ t_1 & t_2 & t_3 \\ t_1^2 & t_2^2 & t_3^2 \end{bmatrix}$$

is invertible when t_1, t_2, and t_3 are distinct. Now the L_i are independent, and since V has dimension 3, these functionals form a basis for V^*. What is the basis for V, of which this is the dual? Such a basis $\{p_1, p_2, p_3\}$ for V must satisfy

$$L_i(p_j) = \delta_{ij}$$

or

$$p_j(t_i) = \delta_{ij}.$$

These polynomial functions are rather easily seen to be

$$p_1(x) = \frac{(x - t_2)(x - t_3)}{(t_1 - t_2)(t_1 - t_3)}$$

$$p_2(x) = \frac{(x - t_1)(x - t_3)}{(t_2 - t_1)(t_2 - t_3)}$$

$$p_3(x) = \frac{(x - t_1)(x - t_2)}{(t_3 - t_1)(t_3 - t_2)}.$$

The basis $\{p_1, p_2, p_3\}$ for V is interesting, because according to (3-14) we have for each p in V

$$p = p(t_1)p_1 + p(t_2)p_2 + p(t_3)p_3.$$

Thus, if c_1, c_2, and c_3 are any real numbers, there is exactly one polynomial function p over R which has degree at most 2 and satisfies $p(t_j) = c_j$, $j = 1, 2, 3$. This polynomial function is $p = c_1p_1 + c_2p_2 + c_3p_3$.

Now let us discuss the relationship between linear functionals and subspaces. If f is a non-zero linear functional, then the rank of f is 1 because the range of f is a non-zero subspace of the scalar field and must (therefore) be the scalar field. If the underlying space V is finite-dimensional, the rank plus nullity theorem (Theorem 2) tells us that the null space N_f has dimension

$$\dim N_f = \dim V - 1.$$

In a vector space of dimension n, a subspace of dimension $n - 1$ is called a **hyperspace.** Such spaces are sometimes called hyperplanes or subspaces of codimension 1. Is every hyperspace the null space of a linear functional? The answer is easily seen to be yes. It is not much more difficult to show that each d-dimensional subspace of an n-dimensional space is the intersection of the null spaces of $(n - d)$ linear functionals (Theorem 16 below).

Definition. *If* V *is a vector space over the field* F *and* S *is a subset of* V, *the* **annihilator** *of* S *is the set* S^0 *of linear functionals* f *on* V *such that* $f(\alpha) = 0$ *for every* α *in* S.

It should be clear to the reader that S^0 is a subspace of V^*, whether S is a subspace of V or not. If S is the set consisting of the zero vector alone, then $S^0 = V^*$. If $S = V$, then S^0 is the zero subspace of V^*. (This is easy to see when V is finite-dimensional.)

Theorem 16. *Let* V *be a finite-dimensional vector space over the field* F, *and let* W *be a subspace of* V. *Then*

$$\dim W + \dim W^0 = \dim V.$$

Proof. Let k be the dimension of W and $\{\alpha_1, \ldots, \alpha_k\}$ a basis for W. Choose vectors $\alpha_{k+1}, \ldots, \alpha_n$ in V such that $\{\alpha_1, \ldots, \alpha_n\}$ is a basis for V. Let $\{f_1, \ldots, f_n\}$ be the basis for V^* which is dual to this basis for V.

The claim is that $\{f_{k+1}, \ldots, f_n\}$ is a basis for the annihilator W^0. Certainly f_i belongs to W^0 for $i \geq k + 1$, because

$$f_i(\alpha_j) = \delta_{ij}$$

and $\delta_{ij} = 0$ if $i \geq k + 1$ and $j \leq k$; from this it follows that, for $i \geq k + 1$, $f_i(\alpha) = 0$ whenever α is a linear combination of $\alpha_1, \ldots, \alpha_k$. The functionals $f_{k+1}, \ldots, f_n$ are independent, so all we must show is that they span W^0. Suppose f is in V^*. Now

$$f = \sum_{i=1}^{n} f(\alpha_i) f_i$$

so that if f is in W^0 we have $f(\alpha_i) = 0$ for $i \leq k$ and

$$f = \sum_{i=k+1}^{n} f(\alpha_i) f_i.$$

We have shown that if $\dim W = k$ and $\dim V = n$ then $\dim W^0 = n - k$. ∎

Corollary. *If* W *is a* k*-dimensional subspace of an* n*-dimensional vector space* V, *then* W *is the intersection of* (n − k) *hyperspaces in* V.

Proof. This is a corollary of the proof of Theorem 16 rather than its statement. In the notation of the proof, W is exactly the set of vectors α such that $f_i(\alpha) = 0$, $i = k + 1, \ldots, n$. In case $k = n - 1$, W is the null space of f_n. ∎

Corollary. *If* W_1 *and* W_2 *are subspaces of a finite-dimensional vector space, then* $W_1 = W_2$ *if and only if* $W_1^0 = W_2^0$.

Proof. If $W_1 = W_2$, then of course $W_1^0 = W_2^0$. If $W_1 \neq W_2$, then one of the two subspaces contains a vector which is not in the other. Suppose there is a vector α which is in W_2 but not in W_1. By the previous corollaries (or the proof of Theorem 16) there is a linear functional f such that $f(\beta) = 0$ for all β in W, but $f(\alpha) \neq 0$. Then f is in W_1^0 but not in W_2^0 and $W_1^0 \neq W_2^0$. ∎

In the next section we shall give different proofs for these two corollaries. The first corollary says that, if we select some ordered basis for the space, each k-dimensional subspace can be described by specifying $(n - k)$ homogeneous linear conditions on the coordinates relative to that basis.

Let us look briefly at systems of homogeneous linear equations from the point of view of linear functionals. Suppose we have a system of linear equations,

$$\begin{aligned} A_{11}x_1 + \cdots + A_{1n}x_n &= 0 \\ \vdots \qquad\qquad\quad & \quad \vdots \\ A_{m1}x_1 + \cdots + A_{mn}x_n &= 0 \end{aligned}$$

for which we wish to find the solutions. If we let f_i, $i = 1, \ldots, m$, be the linear functional on F^n defined by

$$f_i(x_1, \ldots, x_n) = A_{i1}x_1 + \cdots + A_{in}x_n$$

then we are seeking the subspace of F^n of all α such that

$$f_i(\alpha) = 0, \qquad i = 1, \ldots, m.$$

In other words, we are seeking the subspace annihilated by $f_1, \ldots, f_m$. Row-reduction of the coefficient matrix provides us with a systematic method of finding this subspace. The n-tuple $(A_{i1}, \ldots, A_{in})$ gives the coordinates of the linear functional f_i relative to the basis which is dual to the standard basis for F^n. The row space of the coefficient matrix may thus be regarded as the space of linear functionals spanned by $f_1, \ldots, f_m$. The solution space is the subspace annihilated by this space of functionals.

Now one may look at the system of equations from the 'dual' point of view. That is, suppose that we are given m vectors in F^n

$$\alpha_i = (A_{i1}, \ldots, A_{in})$$

and we wish to find the annihilator of the subspace spanned by these vectors. Since a typical linear functional on F^n has the form

$$f(x_1, \ldots, x_n) = c_1x_1 + \cdots + c_nx_n$$

the condition that f be in this annihilator is that

$$\sum_{j=1}^{n} A_{ij}c_j = 0, \qquad i = 1, \ldots, m$$

that is, that $(c_1, \ldots, c_n)$ be a solution of the system $AX = 0$. From this point of view, row-reduction gives us a systematic method of finding the annihilator of the subspace spanned by a given finite set of vectors in F^n.

EXAMPLE 23. Here are three linear functionals on R^4:

$$\begin{aligned} f_1(x_1, x_2, x_3, x_4) &= x_1 + 2x_2 + 2x_3 + x_4 \\ f_2(x_1, x_2, x_3, x_4) &= 2x_2 + x_4 \\ f_3(x_1, x_2, x_3, x_4) &= -2x_1 - 4x_3 + 3x_4. \end{aligned}$$

The subspace which they annihilate may be found explicitly by finding the row-reduced echelon form of the matrix

$$A = \begin{bmatrix} 1 & 2 & 2 & 1 \\ 0 & 2 & 0 & 1 \\ -2 & 0 & -4 & 3 \end{bmatrix}.$$

A short calculation, or a peek at Example 21 of Chapter 2, shows that

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

Therefore, the linear functionals

$$\begin{aligned} g_1(x_1, x_2, x_3, x_4) &= x_1 + 2x_3 \\ g_2(x_1, x_2, x_3, x_4) &= x_2 \\ g_3(x_1, x_2, x_3, x_4) &= x_4 \end{aligned}$$

span the same subspace of $(R^4)^*$ and annihilate the same subspace of R^4 as do f_1, f_2, f_3. The subspace annihilated consists of the vectors with

$$\begin{aligned} x_1 &= -2x_3 \\ x_2 &= x_4 = 0. \end{aligned}$$

EXAMPLE 24. Let W be the subspace of R^5 which is spanned by the vectors

$$\begin{aligned} \alpha_1 &= (2, -2, 3, 4, -1), & \alpha_3 &= (0, 0, -1, -2, 3) \\ \alpha_2 &= (-1, 1, 2, 5, 2), & \alpha_4 &= (1, -1, 2, 3, 0). \end{aligned}$$

How does one describe W^0, the annihilator of W? Let us form the 4×5 matrix A with row vectors $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, and find the row-reduced echelon matrix R which is row-equivalent to A:

$$A = \begin{bmatrix} 2 & -2 & 3 & 4 & -1 \\ -1 & 1 & 2 & 5 & 2 \\ 0 & 0 & -1 & -2 & 3 \\ 1 & -1 & 2 & 3 & 0 \end{bmatrix} \longrightarrow R = \begin{bmatrix} 1 & -1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

If f is a linear functional on R^5:

$$f(x_1, \ldots, x_5) = \sum_{j=1}^{5} c_j x_j$$

then f is in W^0 if and only if $f(\alpha_i) = 0$, $i = 1, 2, 3, 4$, i.e., if and only if

$$\sum_{j=1}^{5} A_{ij} c_j = 0, \qquad 1 \leq i \leq 4.$$

This is equivalent to

$$\sum_{j=1}^{5} R_{ij} c_j = 0, \qquad 1 \leq i \leq 3$$

or

$$\begin{aligned} c_1 - c_2 - c_4 &= 0 \\ c_3 + 2c_4 &= 0 \\ c_5 &= 0. \end{aligned}$$

We obtain all such linear functionals f by assigning arbitrary values to c_2 and c_4, say $c_2 = a$ and $c_4 = b$, and then finding the corresponding $c_1 = a + b$, $c_3 = -2b$, $c_5 = 0$. So W^0 consists of all linear functionals f of the form

$$f(x_1, x_2, x_3, x_4, x_5) = (a + b)x_1 + ax_2 - 2bx_3 + bx_4.$$

The dimension of W^0 is 2 and a basis $\{f_1, f_2\}$ for W^0 can be found by first taking $a = 1$, $b = 0$ and then $a = 0$, $b = 1$:

$$f_1(x_1, \ldots, x_5) = x_1 + x_2$$
$$f_2(x_1, \ldots, x_5) = x_1 - 2x_3 + x_4.$$

The above general f in W^0 is $f = af_1 + bf_2$.

Exercises

1. In R^3, let $\alpha_1 = (1, 0, 1)$, $\alpha_2 = (0, 1, -2)$, $\alpha_3 = (-1, -1, 0)$.

(a) If f is a linear functional on R^3 such that

$$f(\alpha_1) = 1, \qquad f(\alpha_2) = -1, \qquad f(\alpha_3) = 3,$$

and if $\alpha = (a, b, c)$, find $f(\alpha)$.

(b) Describe explicitly a linear functional f on R^3 such that

$$f(\alpha_1) = f(\alpha_2) = 0 \quad \text{but} \quad f(\alpha_3) \neq 0.$$

(c) Let f be any linear functional such that

$$f(\alpha_1) = f(\alpha_2) = 0 \quad \text{and} \quad f(\alpha_3) \neq 0.$$

If $\alpha = (2, 3, -1)$, show that $f(\alpha) \neq 0$.

2. Let $\mathcal{B} = \{\alpha_1, \alpha_2, \alpha_3\}$ be the basis for C^3 defined by

$$\alpha_1 = (1, 0, -1), \qquad \alpha_2 = (1, 1, 1), \qquad \alpha_3 = (2, 2, 0).$$

Find the dual basis of $\mathcal{B}$.

3. If A and B are $n \times n$ matrices over the field F, show that trace (AB) = trace (BA). Now show that similar matrices have the same trace.

4. Let V be the vector space of all polynomial functions p from R into R which have degree 2 or less:

$$p(x) = c_0 + c_1x + c_2x^2.$$

Define three linear functionals on V by

$$f_1(p) = \int_0^1 p(x)\,dx, \qquad f_2(p) = \int_0^2 p(x)\,dx, \qquad f_3(p) = \int_0^{-1} p(x)\,dx.$$

Show that $\{f_1, f_2, f_3\}$ is a basis for V^* by exhibiting the basis for V of which it is the dual.

5. If A and B are $n \times n$ complex matrices, show that $AB - BA = I$ is impossible.

6. Let m and n be positive integers and F a field. Let $f_1, \ldots, f_m$ be linear functionals on F^n. For α in F^n define

$$T\alpha = (f_1(\alpha), \ldots, f_m(\alpha)).$$

Show that T is a linear transformation from F^n into F^m. Then show that every linear transformation from F^n into F^m is of the above form, for some $f_1, \ldots, f_m$.

7. Let $\alpha_1 = (1, 0, -1, 2)$ and $\alpha_2 = (2, 3, 1, 1)$, and let W be the subspace of R^4 spanned by α_1 and α_2. Which linear functionals f:

$$f(x_1, x_2, x_3, x_4) = c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4$$

are in the annihilator of W?

8. Let W be the subspace of R^5 which is spanned by the vectors

$$\alpha_1 = \epsilon_1 + 2\epsilon_2 + \epsilon_3, \qquad \alpha_2 = \epsilon_2 + 3\epsilon_3 + 3\epsilon_4 + \epsilon_5$$
$$\alpha_3 = \epsilon_1 + 4\epsilon_2 + 6\epsilon_3 + 4\epsilon_4 + \epsilon_5.$$

Find a basis for W^0.

9. Let V be the vector space of all 2×2 matrices over the field of real numbers, and let

$$B = \begin{bmatrix} 2 & -2 \\ -1 & 1 \end{bmatrix}.$$

Let W be the subspace of V consisting of all A such that $AB = 0$. Let f be a linear functional on V which is in the annihilator of W. Suppose that $f(I) = 0$ and $f(C) = 3$, where I is the 2×2 identity matrix and

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

Find $f(B)$.

10. Let F be a subfield of the complex numbers. We define n linear functionals on F^n $(n \geq 2)$ by

$$f_k(x_1, \ldots, x_n) = \sum_{j=1}^{n} (k - j)x_j, \qquad 1 \leq k \leq n.$$

What is the dimension of the subspace annihilated by $f_1, \ldots, f_n$?

11. Let W_1 and W_2 be subspaces of a finite-dimensional vector space V.
(a) Prove that $(W_1 + W_2)^0 = W_1^0 \cap W_2^0$.
(b) Prove that $(W_1 \cap W_2)^0 = W_1^0 + W_2^0$.

12. Let V be a finite-dimensional vector space over the field F and let W be a subspace of V. If f is a linear functional on W, prove that there is a linear functional g on V such that $g(\alpha) = f(\alpha)$ for each α in the subspace W.

13. Let F be a subfield of the field of complex numbers and let V be any vector space over F. Suppose that f and g are linear functionals on V such that the function h defined by $h(\alpha) = f(\alpha)g(\alpha)$ is also a linear functional on V. Prove that either $f = 0$ or $g = 0$.

14. Let F be a field of characteristic zero and let V be a finite-dimensional vector space over F. If $\alpha_1, \ldots, \alpha_m$ are finitely many vectors in V, each different from the zero vector, prove that there is a linear functional f on V such that

$$f(\alpha_i) \neq 0, \qquad i = 1, \ldots, m.$$

15. According to Exercise 3, similar matrices have the same trace. Thus we can define the trace of a linear operator on a finite-dimensional space to be the trace of any matrix which represents the operator in an ordered basis. This is well-defined since all such representing matrices for one operator are similar.

Now let V be the space of all 2×2 matrices over the field F and let P be a fixed 2×2 matrix. Let T be the linear operator on V defined by $T(A) = PA$. Prove that trace $(T) = 2$ trace (P).

16. Show that the trace functional on $n \times n$ matrices is unique in the following sense. If W is the space of $n \times n$ matrices over the field F and if f is a linear functional on W such that $f(AB) = f(BA)$ for each A and B in W, then f is a scalar multiple of the trace function. If, in addition, $f(I) = n$, then f is the trace function.

17. Let W be the space of $n \times n$ matrices over the field F, and let W_0 be the subspace spanned by the matrices C of the form $C = AB - BA$. Prove that W_0 is exactly the subspace of matrices which have trace zero. (*Hint:* What is the dimension of the space of matrices of trace zero? Use the matrix 'units,' i.e., matrices with exactly one non-zero entry, to construct enough linearly independent matrices of the form $AB - BA$.)

3.6. The Double Dual

One question about dual bases which we did not answer in the last section was whether every basis for V^* is the dual of some basis for V. One way to answer that question is to consider V^{**}, the dual space of V^*.

If α is a vector in V, then α induces a linear functional L_α on V^* defined by

$$L_\alpha(f) = f(\alpha), \qquad f \text{ in } V^*.$$

The fact that L_α is linear is just a reformulation of the definition of linear operations in V^*:

$$\begin{aligned} L_\alpha(cf + g) &= (cf + g)(\alpha) \\ &= (cf)(\alpha) + g(\alpha) \\ &= cf(\alpha) + g(\alpha) \\ &= cL_\alpha(f) + L_\alpha(g). \end{aligned}$$

If V is finite-dimensional and $\alpha \neq 0$, then $L_\alpha \neq 0$; in other words, there exists a linear functional f such that $f(\alpha) \neq 0$. The proof is very simple and was given in Section 3.5: Choose an ordered basis $\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}$ for V such that $\alpha_1 = \alpha$ and let f be the linear functional which assigns to each vector in V its first coordinate in the ordered basis $\mathcal{B}$.

Theorem 17. *Let* V *be a finite-dimensional vector space over the field* F. *For each vector* α *in* V *define*

$$\mathrm{L}_\alpha(\mathrm{f}) = \mathrm{f}(\alpha), \qquad \mathrm{f} \text{ in } \mathrm{V}^*.$$

The mapping $\alpha \to \mathrm{L}_\alpha$ *is then an isomorphism of* V *onto* V^{**}.

Proof. We showed that for each α the function L_α is linear. Suppose α and β are in V and c is in F, and let $\gamma = c\alpha + \beta$. Then for each f in V^*

$$\begin{aligned} L_\gamma(f) &= f(\gamma) \\ &= f(c\alpha + \beta) \\ &= cf(\alpha) + f(\beta) \\ &= cL_\alpha(f) + L_\beta(f) \end{aligned}$$

and so

$$L_\gamma = cL_\alpha + L_\beta.$$

This shows that the mapping $\alpha \to L_\alpha$ is a linear transformation from V into V^{**}. This transformation is non-singular; for, according to the remarks above $L_\alpha = 0$ if and only if $\alpha = 0$. Now $\alpha \to L_\alpha$ is a non-singular linear transformation from V into V^{**}, and since

$$\dim V^{**} = \dim V^* = \dim V$$

Theorem 9 tells us that this transformation is invertible, and is therefore an isomorphism of V onto V^{**}. ▌

Corollary. *Let* V *be a finite-dimensional vector space over the field* F. *If* L *is a linear functional on the dual space* V* *of* V, *then there is a unique vector* α *in* V *such that*

$$\mathrm{L}(\mathrm{f}) = \mathrm{f}(\alpha)$$

for every f *in* V*.

Corollary. *Let* V *be a finite-dimensional vector space over the field* F. *Each basis for* V* *is the dual of some basis for* V.

Proof. Let $\mathcal{B}^* = \{f_1, \ldots, f_n\}$ be a basis for V^*. By Theorem 15, there is a basis $\{L_1, \ldots, L_n\}$ for V^{**} such that

$$L_i(f_j) = \delta_{ij}.$$

Using the corollary above, for each i there is a vector α_i in V such that

$$L_i(f) = f(\alpha_i)$$

for every f in V^*, i.e., such that $L_i = L_{\alpha_i}$. It follows immediately that $\{\alpha_1, \ldots, \alpha_n\}$ is a basis for V and that $\mathcal{B}^*$ is the dual of this basis. ▌

In view of Theorem 17, we usually identify α with L_α and say that V 'is' the dual space of V^* or that the spaces V, V^* are naturally in duality with one another. Each is the dual space of the other. In the last corollary we have an illustration of how that can be useful. Here is a further illustration.

If E is a subset of V^*, then the annihilator E^0 is (technically) a subset of V^{**}. If we choose to identify V and V^{**} as in Theorem 17, then E^0 is a subspace of V, namely, the set of all α in V such that $f(\alpha) = 0$ for all f in E. In a corollary of Theorem 16 we noted that each subspace W is determined by its annihilator W^0. How is it determined? The answer is that W is the subspace annihilated by all f in W^0, that is, the intersection of the null spaces of all f's in W^0. In our present notation for annihilators, the answer may be phrased very simply: $W = (W^0)^0$.

Theorem 18. *If* S *is any subset of a finite-dimensional vector space* V, *then* $(\mathrm{S}^0)^0$ *is the subspace spanned by* S.

Proof. Let W be the subspace spanned by S. Clearly $W^0 = S^0$. Therefore, what we are to prove is that $W = W^{00}$. We have given one proof. Here is another. By Theorem 16

$$\dim W + \dim W^0 = \dim V$$
$$\dim W^0 + \dim W^{00} = \dim V^*$$

and since $\dim V = \dim V^*$ we have

$$\dim W = \dim W^{00}.$$

Since W is a subspace of W^{00}, we see that $W = W^{00}$. ∎

The results of this section hold for arbitrary vector spaces; however, the proofs require the use of the so-called Axiom of Choice. We want to avoid becoming embroiled in a lengthy discussion of that axiom, so we shall not tackle annihilators for general vector spaces. But, there are two results about linear functionals on arbitrary vector spaces which are so fundamental that we should include them.

Let V be a vector space. We want to define hyperspaces in V. Unless V is finite-dimensional, we cannot do that with the dimension of the hyperspace. But, we can express the idea that a space N falls just one dimension short of filling out V, in the following way:

1. N is a proper subspace of V;
2. if W is a subspace of V which contains N, then either $W = N$ or $W = V$.

Conditions (1) and (2) together say that N is a proper subspace and there is no larger proper subspace, in short, N is a maximal proper subspace.

Definition. *If* V *is a vector space, a* **hyperspace** *in* V *is a maximal proper subspace of* V.

Theorem 19. *If* f *is a non-zero linear functional on the vector space* V, *then the null space of* f *is a hyperspace in* V. *Conversely, every hyperspace in* V *is the null space of a (not unique) non-zero linear functional on* V.

Proof. Let f be a non-zero linear functional on V and N_f its null space. Let α be a vector in V which is not in N_f, i.e., a vector such that $f(\alpha) \neq 0$. We shall show that every vector in V is in the subspace spanned by N_f and α. That subspace consists of all vectors

$$\gamma + c\alpha, \qquad \gamma \text{ in } N_f,\ c \text{ in } F.$$

Let β be in V. Define

$$c = \frac{f(\beta)}{f(\alpha)}$$

which makes sense because $f(\alpha) \neq 0$. Then the vector $\gamma = \beta - c\alpha$ is in N_f since

$$\begin{aligned} f(\gamma) &= f(\beta - c\alpha) \\ &= f(\beta) - cf(\alpha) \\ &= 0. \end{aligned}$$

So β is in the subspace spanned by N_f and α.

Now let N be a hyperspace in V. Fix some vector α which is not in N. Since N is a maximal proper subspace, the subspace spanned by N and α is the entire space V. Therefore each vector β in V has the form

$$\beta = \gamma + c\alpha, \qquad \gamma \text{ in } N,\ c \text{ in } F.$$

The vector γ and the scalar c are uniquely determined by β. If we have also

$$\beta = \gamma' + c'\alpha, \qquad \gamma' \text{ in } N,\ c' \text{ in } F.$$

then

$$(c' - c)\alpha = \gamma - \gamma'.$$

If $c' - c \neq 0$, then α would be in N; hence, $c' = c$ and $\gamma' = \gamma$. Another way to phrase our conclusion is this: If β is in V, there is a unique scalar c such that $\beta - c\alpha$ is in N. Call that scalar $g(\beta)$. It is easy to see that g is a linear functional on V and that N is the null space of g. ∎

Lemma. *If* f *and* g *are linear functionals on a vector space* V, *then* g *is a scalar multiple of* f *if and only if the null space of* g *contains the null space of* f, *that is, if and only if* $f(\alpha) = 0$ *implies* $g(\alpha) = 0$.

Proof. If $f = 0$ then $g = 0$ as well and g is trivially a scalar multiple of f. Suppose $f \neq 0$ so that the null space N_f is a hyperspace in V. Choose some vector α in V with $f(\alpha) \neq 0$ and let

$$c = \frac{g(\alpha)}{f(\alpha)}.$$

The linear functional $h = g - cf$ is 0 on N_f, since both f and g are 0 there, and $h(\alpha) = g(\alpha) - cf(\alpha) = 0$. Thus h is 0 on the subspace spanned by N_f and α—and that subspace is V. We conclude that $h = 0$, i.e., that $g = cf$. ∎

Theorem 20. *Let* $g, f_1, \ldots, f_r$ *be linear functionals on a vector space* V *with respective null spaces* $N, N_1, \ldots, N_r$. *Then* g *is a linear combination of* $f_1, \ldots, f_r$ *if and only if* N *contains the intersection* $N_1 \cap \cdots \cap N_r$.

Proof. If $g = c_1f_1 + \cdots + c_rf_r$ and $f_i(\alpha) = 0$ for each i, then clearly $g(\alpha) = 0$. Therefore, N contains $N_1 \cap \cdots \cap N_r$.

We shall prove the converse (the 'if' half of the theorem) by induction on the number r. The preceding lemma handles the case $r = 1$. Suppose we know the result for $r = k - 1$, and let $f_1, \ldots, f_k$ be linear functionals with null spaces $N_1, \ldots, N_k$ such that $N_1 \cap \cdots \cap N_k$ is contained in N, the

null space of g. Let $g', f_1', \ldots, f_{k-1}'$ be the restrictions of $g, f_1, \ldots, f_{k-1}$ to the subspace N_k. Then $g', f_1', \ldots, f_{k-1}'$ are linear functionals on the vector space N_k. Furthermore, if α is a vector in N_k and $f_i'(\alpha) = 0$, $i = 1, \ldots, k-1$, then α is in $N_1 \cap \cdots \cap N_k$ and so $g'(\alpha) = 0$. By the induction hypothesis (the case $r = k - 1$), there are scalars c_i such that

$$g' = c_1 f_1' + \cdots + c_{k-1} f_{k-1}'.$$

Now let

$$h = g - \sum_{i=1}^{k-1} c_i f_i. \tag{3-16}$$

Then h is a linear functional on V and (3-16) tells us that $h(\alpha) = 0$ for every α in N_k. By the preceding lemma, h is a scalar multiple of f_k. If $h = c_k f_k$, then

$$g = \sum_{i=1}^{k} c_i f_i. \quad \blacksquare$$

Exercises

1. Let n be a positive integer and F a field. Let W be the set of all vectors $(x_1, \ldots, x_n)$ in F^n such that $x_1 + \cdots + x_n = 0$.

(a) Prove that W^0 consists of all linear functionals f of the form

$$f(x_1, \ldots, x_n) = c \sum_{j=1}^{n} x_j.$$

(b) Show that the dual space W^* of W can be 'naturally' identified with the linear functionals

$$f(x_1, \ldots, x_n) = c_1 x_1 + \cdots + c_n x_n$$

on F^n which satisfy $c_1 + \cdots + c_n = 0$.

2. Use Theorem 20 to prove the following. If W is a subspace of a finite-dimensional vector space V and if $\{g_1, \ldots, g_r\}$ is any basis for W^0, then

$$W = \bigcap_{i=1}^{r} N_{g_i}.$$

3. Let S be a set, F a field, and $V(S; F)$ the space of all functions from S into F:

$$(f + g)(x) = f(x) + g(x)$$
$$(cf)(x) = cf(x).$$

Let W be any n-dimensional subspace of $V(S; F)$. Show that there exist points $x_1, \ldots, x_n$ in S and functions $f_1, \ldots, f_n$ in W such that $f_i(x_j) = \delta_{ij}$.

3.7. The Transpose of a Linear Transformation

Suppose that we have two vector spaces over the field F, V, and W, and a linear transformation T from V into W. Then T induces a linear

transformation from W^* into V^*, as follows. Suppose g is a linear functional on W, and let

$$f(\alpha) = g(T\alpha) \tag{3-17}$$

for each α in V. Then (3-17) defines a function f from V into F, namely, the composition of T, a function from V into W, with g, a function from W into F. Since both T and g are linear, Theorem 6 tells us that f is also linear, i.e., f is a linear functional on V. Thus T provides us with a rule T^t which associates with each linear functional g on W a linear functional $f = T^t g$ on V, defined by (3-17). Note also that T^t is actually a linear transformation from W^* into V^*; for, if g_1 and g_2 are in W^* and c is a scalar

$$\begin{aligned}[T^t(cg_1 + g_2)](\alpha) &= (cg_1 + g_2)(T\alpha) \\ &= cg_1(T\alpha) + g_2(T\alpha) \\ &= c(T^t g_1)(\alpha) + (T^t g_2)(\alpha)\end{aligned}$$

so that $T^t(cg_1 + g_2) = cT^t g_1 + T^t g_2$. Let us summarize.

Theorem 21. *Let* V *and* W *be vector spaces over the field* F. *For each linear transformation* T *from* V *into* W, *there is a unique linear transformation* T^t *from* W^* *into* V^* *such that*

$$(T^t g)(\alpha) = g(T\alpha)$$

for every g *in* W^* *and* α *in* V.

We shall call T^t the **transpose** of T. This transformation T^t is often called the adjoint of T; however, we shall not use this terminology.

Theorem 22. *Let* V *and* W *be vector spaces over the field* F, *and let* T *be a linear transformation from* V *into* W. *The null space of* T^t *is the annihilator of the range of* T. *If* V *and* W *are finite-dimensional, then*

(i) *rank* (T^t) = *rank* (T)
(ii) *the range of* T^t *is the annihilator of the null space of* T.

Proof. If g is in W^*, then by definition

$$(T^t g)(\alpha) = g(T\alpha)$$

for each α in V. The statement that g is in the null space of T^t means that $g(T\alpha) = 0$ for every α in V. Thus the null space of T^t is precisely the annihilator of the range of T.

Suppose that V and W are finite-dimensional, say dim $V = n$ and dim $W = m$. For (i): Let r be the rank of T, i.e., the dimension of the range of T. By Theorem 16, the annihilator of the range of T then has dimension $(m - r)$. By the first statement of this theorem, the nullity of T^t must be $(m - r)$. But then since T^t is a linear transformation on an m-dimensional space, the rank of T^t is $m - (m - r) = r$, and so T and T^t have the same rank. For (ii): Let N be the null space of T. Every functional in the range

of T^t is in the annihilator of N; for, suppose $f = T^t g$ for some g in W^*; then, if α is in N

$$f(\alpha) = (T^t g)(\alpha) = g(T\alpha) = g(0) = 0.$$

Now the range of T^t is a subspace of the space N^0, and

$$\dim N^0 = n - \dim N = \text{rank}\,(T) = \text{rank}\,(T^t)$$

so that the range of T^t must be exactly N^0. ∎

Theorem 23. *Let* V *and* W *be finite-dimensional vector spaces over the field* F. *Let* $\mathcal{B}$ *be an ordered basis for* V *with dual basis* $\mathcal{B}^*$, *and let* $\mathcal{B}'$ *be an ordered basis for* W *with dual basis* $\mathcal{B}'^*$. *Let* T *be a linear transformation from* V *into* W; *let* A *be the matrix of* T *relative to* $\mathcal{B}$, $\mathcal{B}'$ *and let* B *be the matrix of* T^t *relative to* $\mathcal{B}'^*$, $\mathcal{B}^*$. *Then* $\mathrm{B}_{ij} = \mathrm{A}_{ji}$.

Proof. Let

$$\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}, \qquad \mathcal{B}' = \{\beta_1, \ldots, \beta_m\},$$
$$\mathcal{B}^* = \{f_1, \ldots, f_n\}, \qquad \mathcal{B}'^* = \{g_1, \ldots, g_m\}.$$

By definition,

$$T\alpha_j = \sum_{i=1}^{m} A_{ij}\beta_i, \qquad j = 1, \ldots, n$$

$$T^t g_j = \sum_{i=1}^{n} B_{ij} f_i, \qquad j = 1, \ldots, m.$$

On the other hand,

$$\begin{aligned}(T^t g_j)(\alpha_i) &= g_j(T\alpha_i)\\ &= g_j\left(\sum_{k=1}^{m} A_{ki}\beta_k\right)\\ &= \sum_{k=1}^{m} A_{ki} g_j(\beta_k)\\ &= \sum_{k=1}^{m} A_{ki}\delta_{jk}\\ &= A_{ji}.\end{aligned}$$

For any linear functional f on V

$$f = \sum_{i=1}^{m} f(\alpha_i) f_i.$$

If we apply this formula to the functional $f = T^t g_j$ and use the fact that $(T^t g_j)(\alpha_i) = A_{ji}$, we have

$$T^t g_j = \sum_{i=1}^{n} A_{ji} f_i$$

from which it immediately follows that $B_{ij} = A_{ji}$. ∎

Definition. *If* A *is an* m $\times$ n *matrix over the field* F, *the* **transpose** *of* A *is the* n $\times$ m *matrix* A^t *defined by* $A^t_{ij} = A_{ji}$.

Theorem 23 thus states that if T is a linear transformation from V into W, the matrix of which in some pair of bases is A, then the transpose transformation T^t is represented in the dual pair of bases by the transpose matrix A^t.

Theorem 24. *Let* A *be any* m $\times$ n *matrix over the field* F. *Then the row rank of* A *is equal to the column rank of* A.

Proof. Let $\mathcal{B}$ be the standard ordered basis for F^n and $\mathcal{B}'$ the standard ordered basis for F^m. Let T be the linear transformation from F^n into F^m such that the matrix of T relative to the pair $\mathcal{B}$, $\mathcal{B}'$ is A, i.e.,

$$T(x_1, \ldots, x_n) = (y_1, \ldots, y_m)$$

where

$$y_i = \sum_{j=1}^{n} A_{ij}x_j.$$

The column rank of A is the rank of the transformation T, because the range of T consists of all m-tuples which are linear combinations of the column vectors of A.

Relative to the dual bases $\mathcal{B}'^*$ and $\mathcal{B}^*$, the transpose mapping T^t is represented by the matrix A^t. Since the columns of A^t are the rows of A, we see by the same reasoning that the row rank of A (the column rank of A^t) is equal to the rank of T^t. By Theorem 22, T and T^t have the same rank, and hence the row rank of A is equal to the column rank of A. ∎

Now we see that if A is an $m \times n$ matrix over F and T is the linear transformation from F^n into F^m defined above, then

$$\text{rank } (T) = \text{row rank } (A) = \text{column rank } (A)$$

and we shall call this number simply the **rank** of A.

EXAMPLE 25. This example will be of a general nature—more discussion than example. Let V be an n-dimensional vector space over the field F, and let T be a linear operator on V. Suppose $\mathcal{B} = \{\alpha_1, \ldots, \alpha_n\}$ is an ordered basis for V. The matrix of T in the ordered basis $\mathcal{B}$ is defined to be the $n \times n$ matrix A such that

$$T\alpha_j = \sum_{j=1}^{n} A_{ij}\alpha_i$$

in other words, A_{ij} is the ith coordinate of the vector $T\alpha_j$ in the ordered basis $\mathcal{B}$. If $\{f_1, \ldots, f_n\}$ is the dual basis of $\mathcal{B}$, this can be stated simply

$$A_{ij} = f_i(T\alpha_j).$$

Let us see what happens when we change basis. Suppose

$$\mathcal{B}' = \{\alpha_1', \ldots, \alpha_n'\}$$

is another ordered basis for V, with dual basis $\{f_1', \ldots, f_n'\}$. If B is the matrix of T in the ordered basis $\mathcal{B}'$, then

$$B_{ij} = f_i'(T\alpha_j').$$

Let U be the invertible linear operator such that $U\alpha_j = \alpha_j'$. Then the transpose of U is given by $U^t f_i' = f_i$. It is easy to verify that since U is invertible, so is U^t and $(U^t)^{-1} = (U^{-1})^t$. Thus $f_i' = (U^{-1})^t f_i$, $i = 1, \ldots, n$. Therefore,

$$\begin{aligned} B_{ij} &= [(U^{-1})^t f_i](T\alpha_j') \\ &= f_i(U^{-1}T\alpha_j') \\ &= f_i(U^{-1}TU\alpha_j). \end{aligned}$$

Now what does this say? Well, $f_i(U^{-1}TU\alpha_j)$ is the i, j entry of the matrix of $U^{-1}TU$ in the ordered basis $\mathcal{B}$. Our computation above shows that this scalar is also the i, j entry of the matrix of T in the ordered basis $\mathcal{B}'$. In other words

$$\begin{aligned} [T]_{\mathcal{B}'} &= [U^{-1}TU]_{\mathcal{B}} \\ &= [U^{-1}]_{\mathcal{B}}[T]_{\mathcal{B}}[U]_{\mathcal{B}} \\ &= [U]_{\mathcal{B}}^{-1}[T]_{\mathcal{B}}[U]_{\mathcal{B}} \end{aligned}$$

and this is precisely the change-of-basis formula which we derived earlier.

Exercises

1. Let F be a field and let f be the linear functional on F^2 defined by $f(x_1, x_2) = ax_1 + bx_2$. For each of the following linear operators T, let $g = T^t f$, and find $g(x_1, x_2)$.

(a) $T(x_1, x_2) = (x_1, 0)$;
(b) $T(x_1, x_2) = (-x_2, x_1)$;
(c) $T(x_1, x_2) = (x_1 - x_2, x_1 + x_2)$.

2. Let V be the vector space of all polynomial functions over the field of real numbers. Let a and b be fixed real numbers and let f be the linear functional on V defined by

$$f(p) = \int_a^b p(x)\, dx.$$

If D is the differentiation operator on V, what is $D^t f$?

3. Let V be the space of all $n \times n$ matrices over a field F and let B be a fixed $n \times n$ matrix. If T is the linear operator on V defined by $T(A) = AB - BA$, and if f is the trace function, what is $T^t f$?

4. Let V be a finite-dimensional vector space over the field F and let T be a linear operator on V. Let c be a scalar and suppose there is a non-zero vector α in V such that $T\alpha = c\alpha$. Prove that there is a non-zero linear functional f on V such that $T^t f = cf$.

5. Let A be an $m \times n$ matrix with *real* entries. Prove that $A = 0$ if and only if trace $(A^tA) = 0$.

6. Let n be a positive integer and let V be the space of all polynomial functions over the field of real numbers which have degree at most n, i.e., functions of the form

$$f(x) = c_0 + c_1x + \cdots + c_nx^n.$$

Let D be the differentiation operator on V. Find a basis for the null space of the transpose operator D^t.

7. Let V be a finite-dimensional vector space over the field F. Show that $T \to T^t$ is an isomorphism of $L(V, V)$ onto $L(V^*, V^*)$.

8. Let V be the vector space of $n \times n$ matrices over the field F.

(a) If B is a fixed $n \times n$ matrix, define a function f_B on V by $f_B(A) =$ trace (B^tA). Show that f_B is a linear functional on V.

(b) Show that every linear functional on V is of the above form, i.e., is f_B for some B.

(c) Show that $B \to f_B$ is an isomorphism of V onto V^*.

CHAPTER 20

Numeric Linear Algebra

This chapter deals with two main topics. The first topic is how to solve linear systems of equations numerically. We start with Gauss elimination, which may be familiar to some readers, but this time in an algorithmic setting with partial pivoting. Variants of this method (Doolittle, Crout, Cholesky, Gauss–Jordan) are discussed in Sec. 20.2. All these methods are direct methods, that is, methods of numerics where we know in advance how many steps they will take until they arrive at a solution. However, small pivots and roundoff error magnification may produce nonsensical results, such as in the Gauss method. A shift occurs in Sec. 20.3, where we discuss numeric iteration methods or indirect methods to address our first topic. Here we cannot be totally sure how many steps will be needed to arrive at a good answer. Several factors—such as how far is the starting value from our initial solution, how is the problem structure influencing speed of convergence, how accurate would we like our result to be—determine the outcome of these methods. Moreover, our computation cycle may not converge. Gauss–Seidel iteration and Jacobi iteration are discussed in Sec. 20.3. Section 20.4 is at the heart of addressing the pitfalls of numeric linear algebra. It is concerned with problems that are ill-conditioned. We learn to estimate how "bad" such a problem is by calculating the condition number of its matrix.

The second topic (Secs. 20.6–20.9) is how to solve eigenvalue problems numerically. Eigenvalue problems appear throughout engineering, physics, mathematics, economics, and many areas. For large or very large matrices, determining the eigenvalues is difficult as it involves finding the roots of the characteristic equations, which are high-degree polynomials. As such, there are different approaches to tackling this problem. Some methods, such as Gerschgorin's method and Collatz's method only provide a range in which eigenvalues lie and thus are known as inclusion methods. Others such as tridiagonalization and QR-factorization actually find all the eigenvalues. The area is quite ingeneous and should be fascinating to the reader.

COMMENT. ***This chapter is independent of Chap. 19 and can be studied immediately after Chap. 7 or 8.***

Prerequisite: Secs. 7.1, 7.2, 8.1.
Sections that may be omitted in a shorter course: 20.4, 20.5, 20.9.
References and Answers to Problems: App. 1 Part E, App. 2.

20.1 Linear Systems: Gauss Elimination

The basic method for solving systems of linear equations by Gauss elimination and back substitution was explained in Sec. 7.3. If you covered Sec. 7.3, you may wonder why we cover Gauss elimination again. The reason is that *here we cover Gauss elimination in the*

www.EngineeringEBooksPdf.com

setting of numerics and introduce new material such as pivoting, row scaling, and operation count. Furthermore, we give an algorithmic representation of Gauss elimination in Table 20.1 that can be readily converted into software. We also show when Gauss elimination runs into difficulties with small pivots and what to do about it. The reader should pay close attention to the material as variants of Gauss elimination are covered in Sec. 20.2 and, furthermore, the general problem of solving linear systems is the focus of the first half of this chapter.

A **linear system of *n* equations** in *n unknowns* $x_1, \cdots, x_n$ is a set of equations $E_1, \cdots, E_n$ of the form

$$
(1) \qquad \begin{aligned}
E_1: &\quad a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\
E_2: &\quad a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\
&\quad \cdots\cdots\cdots\cdots\cdots \\
E_n: &\quad a_{n1}x_1 + \cdots + a_{nn}x_n = b_n
\end{aligned}
$$

where the **coefficients** a_{jk} and the b_j are given numbers. The system is called **homogeneous** if all the b_j are zero; otherwise it is called **nonhomogeneous.** Using matrix multiplication (Sec. 7.2), we can write (1) as a single vector equation

$$
(2) \qquad \mathbf{Ax} = \mathbf{b}
$$

where the **coefficient matrix** $\mathbf{A} = [a_{jk}]$ is the $n \times n$ matrix

$$
\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}
$$

are column vectors. The following matrix $\tilde{\mathbf{A}}$ is called the **augmented matrix** of the system (1):

$$
\tilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{b}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} & b_1 \\ a_{21} & \cdots & a_{2n} & b_2 \\ \cdot & \cdots & \cdot & \cdot \\ a_{n1} & \cdots & a_{nn} & b_n \end{bmatrix}.
$$

A **solution** of (1) is a set of numbers $x_1, \cdots, x_n$ that satisfy all the n equations, and a **solution vector** of (1) is a vector $\mathbf{x}$ whose components constitute a solution of (1).

The method of solving such a system by determinants (Cramer's rule in Sec. 7.7) is not practical, even with efficient methods for evaluating the determinants.

A practical method for the solution of a linear system is the so-called *Gauss elimination*, which we shall now discuss (***proceeding independently of Sec. 7.3***).

www.EngineeringEBooksPdf.com

Gauss Elimination

This standard method for solving linear systems (1) is a systematic process of elimination that reduces (1) to **triangular form** because the system can then be easily solved by **back substitution**. For instance, a triangular system is

$$\begin{aligned} 3x_1 + 5x_2 + 2x_3 &= 8 \\ 8x_2 + 2x_3 &= -7 \\ 6x_3 &= 3 \end{aligned}$$

and back substitution gives $x_3 = \frac{3}{6} = \frac{1}{2}$ from the third equation, then

$$x_2 = \tfrac{1}{8}(-7 - 2x_3) = -1$$

from the second equation, and finally from the first equation

$$x_1 = \tfrac{1}{3}(8 - 5x_2 - 2x_3) = 4.$$

How do we reduce a given system (1) to triangular form? In the first step we ***eliminate*** x_1 from equations E_2 to E_n in (1). We do this by adding (or subtracting) suitable multiples of E_1 to (from) equations $E_2, \cdots, E_n$ and taking the resulting equations, call them $E_2^*, \cdots, E_n^*$ as the new equations. The first equation, E_1, is called the **pivot equation** in this step, and a_{11} is called the **pivot**. This equation is left unaltered. In the second step we take the new second equation E_2^* (which no longer contains x_1) as the pivot equation and use it to ***eliminate*** $\boldsymbol{x_2}$ from E_3^* to E_n^*. And so on. After $n - 1$ steps this gives a triangular system that can be solved by back substitution as just shown. In this way we obtain precisely all solutions of the *given* system (as proved in Sec. 7.3).

The pivot a_{kk} (in step k) ***must be*** different from zero and ***should be*** large in absolute value to avoid roundoff magnification by the multiplication in the elimination. For this we choose as our pivot equation one that has the absolutely largest a_{jk} in column k on or below the main diagonal (actually, the uppermost if there are several such equations). This popular method is called **partial pivoting**. It is used in CASs (e.g., in Maple).

Partial pivoting distinguishes it from **total pivoting**, which involves both row and column interchanges but is hardly used in practice.

Let us illustrate this method with a simple example.

EXAMPLE 1 Gauss Elimination. Partial Pivoting

Solve the system

$$\begin{aligned} E_1&: \quad 8x_2 + 2x_3 = -7 \\ E_2&: \quad 3x_1 + 5x_2 + 2x_3 = 8 \\ E_3&: \quad 6x_1 + 2x_2 + 8x_3 = 26. \end{aligned}$$

Solution. We must pivot since E_1 has no x_1-term. In Column 1, equation E_3 has the largest coefficient. Hence we interchange E_1 and E_3,

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= 26 \\ 3x_1 + 5x_2 + 2x_3 &= 8 \\ 8x_2 + 2x_3 &= -7. \end{aligned}$$

www.EngineeringEBooksPdf.com

Step 1. Elimination of x_1

It would suffice to show the augmented matrix and operate on it. We show both the equations and the augmented matrix. In the first step, the first equation is the pivot equation. Thus

$$\begin{array}{rl} \text{Pivot 6} \longrightarrow & 6x_1 + 2x_2 + 8x_3 = 26 \\ \text{Eliminate} \longrightarrow & \boxed{3x_1} + 5x_2 + 2x_3 = 8 \\ & 8x_2 + 2x_3 = -7 \end{array} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 3 & 5 & 2 & 8 \\ 0 & 8 & 2 & -7 \end{array}\right].$$

To eliminate x_1 from the other equations (here, from the second equation), do:

Subtract $\frac{3}{6} = \frac{1}{2}$ times the pivot equation from the second equation.

The result is

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= 26 \\ 4x_2 - 2x_3 &= -5 \\ 8x_2 + 2x_3 &= -7 \end{aligned} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 0 & 4 & -2 & -5 \\ 0 & 8 & 2 & -7 \end{array}\right].$$

Step 2. Elimination of x_2

The largest coefficient in Column 2 is 8. Hence we take the *new* third equation as the pivot equation, interchanging equations 2 and 3,

$$\begin{array}{rl} & 6x_1 + 2x_2 + 8x_3 = 26 \\ \text{Pivot 8} \longrightarrow & 8x_2 + 2x_3 = -7 \\ \text{Eliminate} \longrightarrow & \boxed{4x_2} - 2x_3 = -5 \end{array} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 0 & 8 & 2 & -7 \\ 0 & 4 & -2 & -5 \end{array}\right].$$

To eliminate x_2 from the third equation, do:

Subtract $\frac{1}{2}$ times the pivot equation from the third equation.

The resulting triangular system is shown below. This is the end of the forward elimination. Now comes the back substitution.

Back substitution. ***Determination of x_3, x_2, x_1***

The triangular system obtained in Step 2 is

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= 26 \\ 8x_2 + 2x_3 &= -7 \\ -3x_3 &= -\tfrac{3}{2} \end{aligned} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 0 & 8 & 2 & -7 \\ 0 & 0 & -3 & -\frac{3}{2} \end{array}\right].$$

From this system, taking the last equation, then the second equation, and finally the first equation, we compute the solution

$$\begin{aligned} x_3 &= \tfrac{1}{2} \\ x_2 &= \tfrac{1}{8}(-7 - 2x_3) = -1 \\ x_1 &= \tfrac{1}{6}(26 - 2x_2 - 8x_3) = 4. \end{aligned}$$

This agrees with the values given above, before the beginning of the example. ■

The general algorithm for the Gauss elimination is shown in Table 20.1. To help explain the algorithm, we have numbered some of its lines. b_j is denoted by $a_{j,n+1}$, for uniformity. In lines 1 and 2 we look for a possible pivot. [For $k = 1$ we can always find one; otherwise x_1 would not occur in (1).] In line 2 we do pivoting if necessary, picking an a_{jk} of greatest absolute value (the one with the smallest j if there are several) and interchange the

www.EngineeringEBooksPdf.com

corresponding rows. If $|a_{kk}|$ is greatest, we do no pivoting. m_{jk} in line 4 suggests *multiplier*, since these are the factors by which we have to multiply the pivot equation E_k^* in Step k before subtracting it from an equation E_j^* below E_k^* from which we want to eliminate x_k. Here we have written E_k^* and E_j^* to indicate that after Step 1 these are no longer the equations given in (1), but these underwent a change in each step, as indicated in line 5. Accordingly, a_{jk} etc. in all lines refer to the most recent equations, and $j \geqq k$ in line 1 indicates that we leave untouched all the equations that have served as pivot equations in previous steps. For $p = k$ in line 5 we get 0 on the right, as it should be in the elimination,

$$a_{jk} - m_{jk}a_{kk} = a_{jk} - \frac{a_{jk}}{a_{kk}}a_{kk} = 0.$$

In line 3, if the last equation in the *triangular* system is $0 = b_n^* \neq 0$, we have no solution. If it is $0 = b_n^* = 0$, we have no unique solution because we then have fewer equations than unknowns.

EXAMPLE 2 Gauss Elimination in Table 20.1, Sample Computation

In Example 1 we had $a_{11} = 0$, so that pivoting was necessary. The greatest coefficient in Column 1 was a_{31}. Thus $\tilde{j} = 3$ in line 2, and we interchanged E_1 and E_3. Then in lines 4 and 5 we computed $m_{21} = \frac{3}{6} = \frac{1}{2}$ and

$$a_{22} = 5 - \tfrac{1}{2} \cdot 2 = 4, \qquad a_{23} = 2 - \tfrac{1}{2} \cdot 8 = -2, \qquad a_{24} = 8 - \tfrac{1}{2} \cdot 26 = -5,$$

and then $m_{31} = \frac{0}{6} = 0$, so that the third equation $8x_2 + 2x_3 = -7$ did not change in Step 1. In Step 2 ($k = 2$) we had 8 as the greatest coefficient in Column 2, hence $\tilde{j} = 3$. We interchanged equations 2 and 3, computed $m_{32} = -\frac{4}{8} = -\frac{1}{2}$ in line 5, and the $a_{33} = -2 - \frac{1}{2} \cdot 2 = -3$, $a_{34} = -5 - \frac{1}{2}(-7) = -\frac{3}{2}$. This produced the triangular form used in the back substitution. ■

If $a_{kk} = 0$ in Step k, ***we must pivot***. If $|a_{kk}|$ is small, ***we should pivot*** because of roundoff error magnification that may seriously affect accuracy or even produce nonsensical results.

EXAMPLE 3 Difficulty with Small Pivots

The solution of the system

$$0.0004x_1 + 1.402x_2 = 1.406$$
$$0.4003x_1 - 1.502x_2 = 2.501$$

is $x_1 = 10$, $x_2 = 1$. We solve this system by the Gauss elimination, using four-digit floating-point arithmetic. (4D is for simplicity. Make an 8D-arithmetic example that shows the same.)

(a) Picking the first of the given equations as the pivot equation, we have to multiply this equation by $m = 0.4003/0.0004 = 1001$ and subtract the result from the second equation, obtaining

$$-1405x_2 = -1404.$$

Hence $x_2 = -1404/(-1405) = 0.9993$, and from the first equation, instead of $x_1 = 10$, we get

$$x_1 = \frac{1}{0.0004}(1.406 - 1.402 \cdot 0.9993) = \frac{0.005}{0.0004} = 12.5.$$

This failure occurs because $|a_{11}|$ is small compared with $|a_{12}|$, so that a small roundoff error in x_2 leads to a large error in x_1.

www.EngineeringEBooksPdf.com

(b) Picking the second of the given equations as the pivot equation, we have to multiply this equation by $0.0004/0.4003 = 0.0009993$ and subtract the result from the first equation, obtaining

$$1.404x_2 = 1.404.$$

Hence $x_2 = 1$, and from the pivot equation $x_1 = 10$. This success occurs because $|a_{21}|$ is not very small compared to $|a_{22}|$, so that a small roundoff error in x_2 would not lead to a large error in x_1. Indeed, for instance, if we had the value $x_2 = 1.002$, we would still have from the pivot equation the good value $x_1 = (2.501 + 1.505)/0.4003 = 10.01$. ■

Table 20.1 Gauss Elimination

ALGORITHM GAUSS ($\tilde{\mathbf{A}} = [a_{jk}] = [\mathbf{A} \quad \mathbf{b}]$)

This algorithm computes a unique solution $\mathbf{x} = [x_j]$ of the system (1) or indicates that (1) has no unique solution.

INPUT: Augmented $n \times (n + 1)$ matrix $\tilde{\mathbf{A}} = [a_{jk}]$, where $a_{j,n+1} = b_j$

OUTPUT: Solution $\mathbf{x} = [x_j]$ of (1) or message that the system (1) has no unique solution

For $k = 1, \cdots, n - 1$, do:

1 $\quad m = k$

For $j = k + 1, \cdots, n$, do:

If $(|a_{mk}| < |a_{jk}|)$ then $m = j$

End

If $a_{mk} = 0$ then OUTPUT "No unique solution exists"

Stop

[*Procedure completed unsuccessfully*]

2 $\quad$ Else exchange row k and row m

3 $\quad$ If $a_{nn} = 0$ then OUTPUT "No unique solution exists."

Stop

Else

4 $\quad$ For $j = k + 1, \cdots, n$, do:

$$m_{jk} := \frac{a_{jk}}{a_{kk}}$$

5 $\quad$ For $p = k + 1, \cdots, n + 1$, do:

$$a_{jp} := a_{jp} - m_{jk}a_{kp}$$

End

End

End

6 $$x_n = \frac{a_{n,n+1}}{a_{nn}} \qquad [\textit{Start back substitution}]$$

For $i = n - 1, \cdots, 1$, do:

7 $$x_i = \frac{1}{a_{ii}}\left(a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j\right)$$

End

OUTPUT $\mathbf{x} = [x_j]$. Stop

End GAUSS

www.EngineeringEBooksPdf.com

Error estimates for the Gauss elimination are discussed in Ref. [E5] listed in App. 1.

Row scaling means the multiplication of each Row j by a suitable scaling factor s_j. It is done in connection with partial pivoting to get more accurate solutions. Despite much research (see Refs. [E9], [E24] in App. 1) and the proposition of several principles, scaling is still not well understood. As a possibility, one can scale for pivot choice only (not in the calculation, to avoid additional roundoff) and take as first pivot the entry a_{j1} for which $|a_{j1}|/|A_j|$ is largest; here A_j is an entry of largest absolute value in Row j. Similarly in the further steps of the Gauss elimination.

For instance, for the system

$$4.0000x_1 + 14020x_2 = 14060$$
$$0.4003x_1 - 1.502x_2 = 2.501$$

we might pick 4 as pivot, but dividing the first equation by 10^4 gives the system in Example 3, for which the second equation is a better pivot equation.

Operation Count

Quite generally, important factors in judging the quality of a numeric method are

Amount of storage

Amount of time ($\equiv$ number of operations)

Effect of roundoff error

For the Gauss elimination, the operation count for a full matrix (a matrix with relatively many nonzero entries) is as follows. In Step k we eliminate x_k from $n - k$ equations. This needs $n - k$ divisions in computing the m_{jk} (line 3) and $(n - k)(n - k + 1)$ multiplications and as many subtractions (both in line 4). Since we do $n - 1$ steps, k goes from 1 to $n - 1$ and thus the total number of operations in this forward elimination is

$$f(n) = \sum_{k=1}^{n-1} (n - k) + 2\sum_{k=1}^{n-1} (n - k)(n - k + 1) \qquad \text{(write } n - k = s\text{)}$$
$$= \sum_{s=1}^{n-1} s + 2\sum_{s=1}^{n-1} s(s + 1) = \tfrac{1}{2}(n - 1)n + \tfrac{2}{3}(n^2 - 1)n \approx \tfrac{2}{3}n^3$$

where $2n^3/3$ is obtained by dropping lower powers of n. We see that $f(n)$ grows about proportional to n^3. We say that $f(n)$ is of *order* n^3 and write

$$f(n) = O(n^3)$$

where O suggests **order**. The general definition of O is as follows. We write

$$f(n) = O(h(n))$$

if the quotients $|f(n)/h(n)|$ and $|h(n)/f(n)|$ remain bounded (do not trail off to infinity) as $n \to \infty$. In our present case, $h(n) = n^3$ and, indeed, $f(n)/n^3 \to \frac{2}{3}$ because the omitted terms divided by n^3 go to zero as $n \to \infty$.

www.EngineeringEBooksPdf.com

In the back substitution of x_i we make $n - i$ multiplications and as many subtractions, as well as 1 division. Hence the number of operations in the back substitution is

$$b(n) = 2\sum_{i=1}^{n}(n - i) + n = 2\sum_{s=1}^{n} s + n = n(n + 1) + n = n^2 + 2n = O(n^2).$$

We see that it grows more slowly than the number of operations in the forward elimination of the Gauss algorithm, so that it is negligible for large systems because it is smaller by a factor n, approximately. For instance, if an operation takes 10^{-9} sec, then the times needed are:

Algorithm	$n = 1000$	$n = 10000$
Elimination	0.7 sec	11 min
Back substitution	0.001 sec	0.1 sec

PROBLEM SET 20.1

APPLICATIONS of linear systems see Secs. 7.1 and 8.2.

1–3 GEOMETRIC INTERPRETATION

Solve graphically and explain geometrically.

1. $x_1 - 4x_2 = 20.1$
$3x_1 + 5x_2 = 5.9$

2. $-5.00x_1 + 8.40x_2 = 0$
$10.25x_1 - 17.22x_2 = 0$

3. $7.2x_1 - 3.5x_2 = 16.0$
$-14.4x_1 + 7.0x_2 = 31.0$

4–16 GAUSS ELIMINATION

Solve the following linear systems by Gauss elimination, with partial pivoting if necessary (but without scaling). Show the intermediate steps. Check the result by substitution. If no solution or more than one solution exists, give a reason.

4. $6x_1 + x_2 = -3$
$4x_1 - 2x_2 = 6$

5. $2x_1 - 8x_2 = -4$
$3x_1 + x_2 = 7$

6. $25.38x_1 - 15.48x_2 = 30.60$
$-14.10x_1 + 8.60x_2 = -17.00$

7. $-3x_1 + 6x_2 - 9x_3 = -46.725$
$x_1 - 4x_2 + 3x_3 = 19.571$
$2x_1 + 5x_2 - 7x_3 = -20.073$

8. $5x_1 + 3x_2 + x_3 = 2$
$-4x_2 + 8x_3 = -3$
$10x_1 - 6x_2 + 26x_3 = 0$

9. $6x_2 + 13x_3 = 137.86$
$6x_1 - 8x_3 = -85.88$
$13x_1 - 8x_2 = 178.54$

10. $4x_1 + 4x_2 + 2x_3 = 0$
$3x_1 - x_2 + 2x_3 = 0$
$3x_1 + 7x_2 + x_3 = 0$

11. $3.4x_1 - 6.12x_2 - 2.72x_3 = 0$
$-x_1 + 1.80x_2 + 0.80x_3 = 0$
$2.7x_1 - 4.86x_2 + 2.16x_3 = 0$

12. $5x_1 + 3x_2 + x_3 = 2$
$-4x_2 + 8x_3 = -3$
$10x_1 - 6x_2 + 26x_3 = 0$

www.EngineeringEBooksPdf.com

13.
$$\begin{aligned} 3x_2 + 5x_3 &= 1.20736 \\ 3x_1 - 4x_2 &= -2.34066 \\ 5x_1 + 6x_3 &= -0.329193 \end{aligned}$$

14.
$$\begin{aligned} -47x_1 + 4x_2 - 7x_3 &= -118 \\ 19x_1 - 3x_2 + 2x_3 &= 43 \\ -15x_1 + 5x_2 &= -25 \end{aligned}$$

15.
$$\begin{aligned} 2.2x_2 + 1.5x_3 - 3.3x_4 &= -9.30 \\ 0.2x_1 + 1.8x_2 + 4.2x_4 &= 9.24 \\ -x_1 - 3.1x_2 + 2.5x_3 &= -8.70 \\ 0.5x_1 - 3.8x_3 + 1.5x_4 &= 11.94 \end{aligned}$$

16.
$$\begin{aligned} 3.2x_1 + 1.6x_2 &= -0.8 \\ 1.6x_1 - 0.8x_2 + 2.4x_3 &= 16.0 \\ 2.4x_2 - 4.8x_3 + 3.6x_4 &= -39.0 \\ 3.6x_3 + 2.4x_4 &= 10.2 \end{aligned}$$

17. CAS EXPERIMENT. Gauss Elimination. Write a program for the Gauss elimination with pivoting. Apply it to Probs. 13–16. Experiment with systems whose coefficient determinant is small in absolute value. Also investigate the performance of your program for larger systems of your choice, including sparse systems.

18. TEAM PROJECT. Linear Systems and Gauss Elimination. (a) Existence and uniqueness. Find a and b such that $ax_1 + x_2 = b, x_1 + x_2 = 3$ has (i) a unique solution, (ii) infinitely many solutions, (iii) no solutions.

(b) Gauss elimination and nonexistence. Apply the Gauss elimination to the following two systems and compare the calculations step by step. Explain why the elimination fails if no solution exists.

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ 4x_1 + 2x_2 - x_3 &= 5 \\ 9x_1 + 5x_2 - x_3 &= 13 \end{aligned}$$

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ 4x_1 + 2x_2 - x_3 &= 5 \\ 9x_1 + 5x_2 - x_3 &= 12. \end{aligned}$$

(c) Zero determinant. Why may a computer program give you the result that a homogeneous linear system has only the trivial solution although you know its coefficient determinant to be zero?

(d) Pivoting. Solve System (A) (below) by the Gauss elimination first without pivoting. Show that for any fixed machine word length and sufficiently small $\epsilon > 0$ the computer gives $x_2 = 1$ and then $x_1 = 0$. What is the exact solution? Its limit as $\epsilon \to 0$? Then solve the system by the Gauss elimination with pivoting. Compare and comment.

(e) Pivoting. Solve System (B) by the Gauss elimination and three-digit rounding arithmetic, choosing (i) the first equation, (ii) the second equation as pivot equation. (Remember to round to 3S after each operation before doing the next, just as would be done on a computer!) Then use four-digit rounding arithmetic in those two calculations. Compare and comment.

$$\text{(A)} \qquad \begin{aligned} \epsilon x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned}$$

$$\text{(B)} \qquad \begin{aligned} 4.03x_1 + 2.16x_2 &= -4.61 \\ 6.21x_1 + 3.35x_2 &= -7.19 \end{aligned}$$

20.2 Linear Systems: LU-Factorization, Matrix Inversion

We continue our discussion of numeric methods for solving linear systems of n equations in n unknowns $x_1, \cdots, x_n$,

$$\mathbf{Ax} = \mathbf{b} \tag{1}$$

where $\mathbf{A} = [a_{jk}]$ is the $n \times n$ given coefficient matrix and $\mathbf{x}^\mathsf{T} = [x_1, \cdots, x_n]$ and $\mathbf{b}^\mathsf{T} = [b_1, \cdots, b_n]$. We present three related methods that are modifications of the Gauss

www.EngineeringEBooksPdf.com

elimination, which require fewer arithmetic operations. They are named after Doolittle, Crout, and Cholesky and use the idea of the LU-factorization of **A**, which we explain first.

An **LU-factorization** of a given square matrix **A** is of the form

$$\mathbf{A} = \mathbf{LU} \tag{2}$$

where **L** is ***lower triangular*** and **U** is ***upper triangular***. For example,

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 8 & 5 \end{bmatrix} = \mathbf{LU} = \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 0 & -7 \end{bmatrix}.$$

It can be proved that for any nonsingular matrix (see Sec. 7.8) the rows can be reordered so that the resulting matrix **A** has an LU-factorization (2) in which **L** turns out to be the matrix of the ***multipliers*** m_{jk} of the Gauss elimination, with main diagonal $1, \cdots, 1$, and **U** is the matrix of the triangular system at the end of the Gauss elimination. (See Ref. [E5], pp. 155–156, listed in App. 1.)

The ***crucial idea*** now is that **L** and **U** in (2) can be computed directly, without solving simultaneous equations (thus, without using the Gauss elimination). As a count shows, this needs about $n^3/3$ operations, about half as many as the Gauss elimination, which needs about $2n^3/3$ (see Sec. 20.1). And once we have (2), we can use it for solving $\mathbf{Ax} = \mathbf{b}$ in two steps, involving only about n^2 operations, simply by noting that $\mathbf{Ax} = \mathbf{LUx} = \mathbf{b}$ may be written

$$\text{(a)} \quad \mathbf{Ly} = \mathbf{b} \qquad \text{where} \qquad \text{(b)} \quad \mathbf{Ux} = \mathbf{y} \tag{3}$$

and solving first (3a) for **y** and then (3b) for **x**. Here we can require that **L** have main diagonal $1, \cdots, 1$ as stated before; then this is called **Doolittle's method**.[1] Both systems (3a) and (3b) are triangular, so we can solve them as in the back substitution for the Gauss elimination.

A similar method, **Crout's method**,[2] is obtained from (2) if **U** (instead of **L**) is required to have main diagonal $1, \cdots, 1$. In either case the factorization (2) is unique.

EXAMPLE 1 Doolittle's Method

Solve the system in Example 1 of Sec. 20.1 by Doolittle's method.

Solution. The decomposition (2) is obtained from

$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 6 & 2 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

[1]MYRICK H. DOOLITTLE (1830–1913). American mathematician employed by the U.S. Coast and Geodetic Survey Office. His method appeared in *U.S. Coast and Geodetic Survey*, 1878, 115–120.

[2]PRESCOTT DURAND CROUT (1907–1984), American mathematician, professor at MIT, also worked at General Electric.

www.EngineeringEBooksPdf.com

by determining the m_{jk} and u_{jk}, using matrix multiplication. By going through **A** row by row we get successively

$a_{11} = 3 = 1 \cdot u_{11} = u_{11}$	$a_{12} = 5 = 1 \cdot u_{12} = u_{12}$	$a_{13} = 2 = 1 \cdot u_{13} = u_{13}$
$a_{21} = 0 = m_{21}u_{11}$ $m_{21} = 0$	$a_{22} = 8 = m_{21}u_{12} + u_{22}$ $u_{22} = 8$	$a_{23} = 2 = m_{21}u_{13} + u_{23}$ $u_{23} = 2$
$a_{31} = 6 = m_{31}u_{11}$ $= m_{31} \cdot 3$ $m_{31} = 2$	$a_{32} = 2 = m_{31}u_{12} + m_{32}u_{22}$ $= 2 \cdot 5 + m_{32} \cdot 8$ $m_{32} = -1$	$a_{33} = 8 = m_{31}u_{13} + m_{32}u_{23} + u_{33}$ $= 2 \cdot 2 - 1 \cdot 2 + u_{33}$ $u_{33} = 6$

Thus the factorization (2) is

$$\begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 6 & 2 & 8 \end{bmatrix} = \mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 0 & 0 & 6 \end{bmatrix}.$$

We first solve $\mathbf{Ly} = \mathbf{b}$, determining $y_1 = 8$, then $y_2 = -7$, then y_3 from $2y_1 - y_2 + y_3 = 16 + 7 + y_3 = 26$; thus (note the interchange in **b** because of the interchange in **A**!)

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 8 \\ -7 \\ 26 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{y} = \begin{bmatrix} 8 \\ -7 \\ 3 \end{bmatrix}.$$

Then we solve $\mathbf{Ux} = \mathbf{y}$, determining $x_3 = \frac{3}{6}$ then x_2, then x_1, that is,

$$\begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8 \\ -7 \\ 3 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{x} = \begin{bmatrix} 4 \\ -1 \\ \frac{1}{2} \end{bmatrix}.$$

This agrees with the solution in Example 1 of Sec. 20.1. ■

Our formulas in Example 1 suggest that for general n the entries of the matrices $\mathbf{L} = [m_{jk}]$ (with main diagonal $1, \cdots, 1$ and m_{jk} suggesting "multiplier") and $\mathbf{U} = [u_{jk}]$ in the **Doolittle method** are computed from

$$\begin{aligned} u_{1k} &= a_{1k} && k = 1, \cdots, n \\ m_{j1} &= \frac{a_{j1}}{u_{11}} && j = 2, \cdots, n \\ u_{jk} &= a_{jk} - \sum_{s=1}^{j-1} m_{js}u_{sk} && k = j, \cdots, n; \quad j \geqq 2 \\ m_{jk} &= \frac{1}{u_{kk}} \left(a_{jk} - \sum_{s=1}^{k-1} m_{js}u_{sk} \right) && j = k+1, \cdots, n; \quad k \geqq 2. \end{aligned} \tag{4}$$

www.EngineeringEBooksPdf.com

Row Interchanges. Matrices, such as

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

have no LU-factorization (try!). This indicates that for obtaining an LU-factorization, row interchanges of **A** (and corresponding interchanges in **b**) may be necessary.

Cholesky's Method

For a ***symmetric, positive definite*** matrix **A** (thus $\mathbf{A} = \mathbf{A}^\mathsf{T}$, $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$) we can in (2) even choose $\mathbf{U} = \mathbf{L}^\mathsf{T}$, thus $u_{jk} = m_{kj}$ (but cannot impose conditions on the main diagonal entries). For example,

$$\mathbf{A} = \begin{bmatrix} 4 & 2 & 14 \\ 2 & 17 & -5 \\ 14 & -5 & 83 \end{bmatrix} = \mathbf{L}\mathbf{L}^\mathsf{T} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 7 & -3 & 5 \end{bmatrix} \begin{bmatrix} 2 & 1 & 7 \\ 0 & 4 & -3 \\ 0 & 0 & 5 \end{bmatrix}. \tag{5}$$

The popular method of solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ based on this factorization $\mathbf{A} = \mathbf{L}\mathbf{L}^\mathsf{T}$ is called **Cholesky's method**.[3] In terms of the entries of $\mathbf{L} = [l_{jk}]$ the formulas for the factorization are

$$\begin{aligned} l_{11} &= \sqrt{a_{11}} \\ l_{j1} &= \frac{a_{j1}}{l_{11}} && j = 2, \cdots, n \\ l_{jj} &= \sqrt{a_{jj} - \sum_{s=1}^{j-1} l_{js}^2} && j = 2, \cdots, n \\ l_{pj} &= \frac{1}{l_{jj}}\left(a_{pj} - \sum_{s=1}^{j-1} l_{js} l_{ps}\right) && p = j+1, \cdots, n; \quad j \geqq 2. \end{aligned} \tag{6}$$

If **A** is symmetric but not positive definite, this method could still be applied, but then leads to a *complex* matrix **L**, so that the method becomes impractical.

EXAMPLE 2 Cholesky's Method

Solve by Cholesky's method:

$$\begin{aligned} 4x_1 + 2x_2 + 14x_3 &= 14 \\ 2x_1 + 17x_2 - 5x_3 &= -101 \\ 14x_1 - 5x_2 + 83x_3 &= 155. \end{aligned}$$

[3]ANDRÉ-LOUIS CHOLESKY (1875–1918), French military officer, geodecist, and mathematician. Surveyed Crete and North Africa. Died in World War I. His method was published posthumously in *Bulletin Géodésique* in 1924 but received little attention until JOHN TODD (1911–2007) — Irish-American mathematician, numerical analysist, and early pioneer of computer methods in numerics, professor at Caltech, and close personal friend and collaborator of ERWIN KREYSZIG, see [E20]—taught Cholesky's method in his analysis course at King's College, London, in the 1940s.

www.EngineeringEBooksPdf.com

Solution. From (6) or from the form of the factorization

$$\begin{bmatrix} 4 & 2 & 14 \\ 2 & 17 & -5 \\ 14 & -5 & 83 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

we compute, in the given order,

$$l_{11} = \sqrt{a_{11}} = 2 \qquad l_{21} = \frac{a_{21}}{l_{11}} = \frac{2}{2} = 1 \qquad l_{31} = \frac{a_{31}}{l_{11}} = \frac{14}{2} = 7$$

$$l_{22} = \sqrt{a_{22} - l_{21}^2} = \sqrt{17 - 1} = 4$$

$$l_{32} = \frac{1}{l_{23}}(a_{32} - l_{31}l_{21}) = \frac{1}{4}(-5 - 7 \cdot 1) = -3$$

$$l_{33} = \sqrt{a_{33} - l_{31}^2 - l_{32}^2} = \sqrt{83 - 7^2 - (-3)^2} = 5.$$

This agrees with (5). We now have to solve $\mathbf{Ly} = \mathbf{b}$, that is,

$$\begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 7 & -3 & 5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 14 \\ -101 \\ 155 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{y} = \begin{bmatrix} 7 \\ -27 \\ 5 \end{bmatrix}.$$

As the second step, we have to solve $\mathbf{Ux} = \mathbf{L}^\mathsf{T}\mathbf{x} = \mathbf{y}$, that is,

$$\begin{bmatrix} 2 & 1 & 7 \\ 0 & 4 & -3 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ -27 \\ 5 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{x} = \begin{bmatrix} 3 \\ -6 \\ 1 \end{bmatrix}.$$

■

THEOREM 1

Stability of the Cholesky Factorization

The Cholesky LL^T-factorization is numerically stable (as defined in Sec. 19.1).

PROOF We have $a_{jj} = l_{j1}^2 + l_{j2}^2 + \cdots + l_{jj}^2$ by squaring the third formula in (6) and solving it for a_{jj}. Hence for all l_{jk} (note that $l_{jk} = 0$ for $k > j$) we obtain (the inequality being trivial)

$$l_{jk}^2 \leqq l_{j1}^2 + l_{j2}^2 + \cdots + l_{jj}^2 = a_{jj}.$$

That is, l_{jk}^2 is bounded by an entry of $\mathbf{A}$, which means stability against rounding. ■

Gauss–Jordan Elimination. Matrix Inversion

Another variant of the Gauss elimination is the **Gauss–Jordan elimination**, introduced by W. Jordan in 1920, in which back substitution is avoided by additional computations that reduce the matrix to diagonal form, instead of the triangular form in the Gauss elimination. But this reduction from the Gauss triangular to the diagonal form requires more operations than back substitution does, so that the method is *disadvantageous* for solving systems $\mathbf{Ax} = \mathbf{b}$. But it may be used for matrix inversion, where the situation is as follows.

www.EngineeringEBooksPdf.com

The **inverse** of a nonsingular square matrix $\mathbf{A}$ may be determined in principle by solving the n systems

$$\mathbf{Ax} = \mathbf{b}_j \qquad (j = 1, \cdots, n) \tag{7}$$

where $\mathbf{b}_j$ is the jth column of the $n \times n$ unit matrix.

However, it is preferable to produce $\mathbf{A}^{-1}$ by operating on the unit matrix $\mathbf{I}$ in the same way as the Gauss–Jordan algorithm, reducing $\mathbf{A}$ to $\mathbf{I}$. A typical illustrative example of this method is given in Sec. 7.8.

PROBLEM SET 20.2

1–5 DOOLITTLE'S METHOD

Show the factorization and solve by Doolittle's method.

1. $4x_1 + 5x_2 = 14$
$12x_1 + 14x_2 = 36$

2. $2x_1 + 9x_2 = 82$
$3x_1 - 5x_2 = -62$

3. $5x_1 + 4x_2 + x_3 = 6.8$
$10x_1 + 9x_2 + 4x_3 = 17.6$
$10x_1 + 13x_2 + 15x_3 = 38.4$

4. $2x_1 + x_2 + 2x_3 = 0$
$-2x_1 + 2x_2 + x_3 = 0$
$x_1 + 2x_2 - 2x_3 = 18$

5. $3x_1 + 9x_2 + 6x_3 = 4.6$
$18x_1 + 48x_2 + 39x_3 = 27.2$
$9x_1 - 27x_2 + 42x_3 = 9.0$

6. TEAM PROJECT. Crout's method factorizes $\mathbf{A} = \mathbf{LU}$, where $\mathbf{L}$ is lower triangular and $\mathbf{U}$ is upper triangular with diagonal entries $u_{jj} = 1, j = 1, \cdots, n$.

(a) Formulas. Obtain formulas for Crout's method similar to (4).

(b) Examples. Solve Prob. 5 by Crout's method.

(c) Factor the following matrix by the Doolittle, Crout, and Cholesky methods.

$$\begin{bmatrix} 1 & -4 & 2 \\ -4 & 25 & 4 \\ 2 & 4 & 24 \end{bmatrix}$$

(d) Give the formulas for factoring a tridiagonal matrix by Crout's method.

(e) When can you obtain Crout's factorization from Doolittle's by transposition?

7–12 CHOLESKY'S METHOD

Show the factorization and solve.

7. $9x_1 + 6x_2 + 12x_3 = 17.4$
$6x_1 + 13x_2 + 11x_3 = 23.6$
$12x_1 + 11x_2 + 26x_3 = 30.8$

8. $4x_1 + 6x_2 + 8x_3 = 0$
$6x_1 + 34x_2 + 52x_3 = -160$
$8x_1 + 52x_2 + 129x_3 = -452$

9. $0.01x_1 + 0.03x_3 = 0.14$
$0.16x_2 + 0.08x_3 = 0.16$
$0.03x_1 + 0.08x_2 + 0.14x_3 = 0.54$

10. $4x_1 + 2x_3 = 1.5$
$4x_2 + x_3 = 4.0$
$2x_1 + x_2 + 2x_3 = 2.5$

11. $x_1 - x_2 + 3x_3 + 2x_4 = 15$
$-x_1 + 5x_2 - 5x_3 - 2x_4 = -35$
$3x_1 - 5x_2 + 19x_3 + 3x_4 = 94$
$2x_1 - 2x_2 + 3x_3 + 21x_4 = 1$

12. $4x_1 + 2x_2 + 4x_3 = 20$
$2x_1 + 2x_2 + 3x_3 + 2x_4 = 36$
$4x_1 + 3x_2 + 6x_3 + 3x_4 = 60$
$2x_2 + 3x_3 + 9x_4 = 122$

13. Definiteness. Let $\mathbf{A}, \mathbf{B}$ be $n \times n$ and positive definite. Are $-\mathbf{A}, \mathbf{A}^{\mathsf{T}}, \mathbf{A} + \mathbf{B}, \mathbf{A} - \mathbf{B}$ positive definite?

www.EngineeringEBooksPdf.com

14. CAS PROJECT. Cholesky's Method. (a) Write a program for solving linear systems by Cholesky's method and apply it to Example 2 in the text, to Probs. 7–9, and to systems of your choice.

(b) Splines. Apply the factorization part of the program to the following matrices (as they occur in (9), Sec. 19.4 (with $c_j = 1$), in connection with splines).

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix}, \quad \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}.$$

15–19 INVERSE

Find the inverse by the Gauss–Jordan method, showing the details.

15. In Prob. 1

16. In Prob. 4

17. In Team Project 6(c)

18. In Prob. 9

19. In Prob. 12

20. Rounding. For the following matrix **A** find det **A**. What happens if you roundoff the given entries to **(a)** 5S, **(b)** 4S, **(c)** 3S, **(d)** 2S, **(e)** 1S? What is the practical implication of your work?

$$\mathbf{A} = \begin{bmatrix} \frac{1}{3} & \frac{1}{4} & 2 \\ -\frac{1}{9} & 1 & \frac{1}{7} \\ \frac{4}{63} & -\frac{3}{28} & \frac{13}{49} \end{bmatrix}$$

20.3 Linear Systems: Solution by Iteration

The Gauss elimination and its variants in the last two sections belong to the **direct methods** for solving linear systems of equations; these are methods that give solutions after an amount of computation that can be specified in advance. In contrast, in an **indirect** or **iterative method** we start from an approximation to the true solution and, if successful, obtain better and better approximations from a computational cycle repeated as often as may be necessary for achieving a required accuracy, so that the amount of arithmetic depends upon the accuracy required and varies from case to case.

We apply iterative methods if the convergence is rapid (if matrices have large main diagonal entries, as we shall see), so that we save operations compared to a direct method. We also use iterative methods if a large system is **sparse**, that is, has very many zero coefficients, so that one would waste space in storing zeros, for instance, 9995 zeros per equation in a potential problem of 10^4 equations in 10^4 unknowns with typically only 5 nonzero terms per equation (more on this in Sec. 21.4).

Gauss–Seidel Iteration Method[4]

This is an iterative method of great practical importance, which we can simply explain in terms of an example.

EXAMPLE 1 Gauss–Seidel Iteration

We consider the linear system

$$\begin{aligned} x_1 - 0.25x_2 - 0.25x_3 \qquad\qquad &= 50 \\ -0.25x_1 + x_2 \qquad\qquad - 0.25x_4 &= 50 \\ -0.25x_1 \qquad\qquad + x_3 - 0.25x_4 &= 25 \\ - 0.25x_2 - 0.25x_3 + x_4 &= 25. \end{aligned} \tag{1}$$

[4]PHILIPP LUDWIG VON SEIDEL (1821–1896), German mathematician. For Gauss see footnote 5 in Sec. 5.4.

www.EngineeringEBooksPdf.com

(Equations of this form arise in the numeric solution of PDEs and in spline interpolation.) We write the system in the form

$$
(2)\qquad
\begin{aligned}
x_1 &= \phantom{0.25x_1 +{}} 0.25x_2 + 0.25x_3 \phantom{{}+ 0.25x_4} + 50\\
x_2 &= 0.25x_1 \phantom{{}+ 0.25x_2} \phantom{{}+ 0.25x_3} + 0.25x_4 + 50\\
x_3 &= 0.25x_1 \phantom{{}+ 0.25x_2} \phantom{{}+ 0.25x_3} + 0.25x_4 + 25\\
x_4 &= \phantom{0.25x_1 +{}} 0.25x_2 + 0.25x_3 \phantom{{}+ 0.25x_4} + 25.
\end{aligned}
$$

These equations are now used for iteration; that is, we start from a (possibly poor) approximation to the solution, say $x_1^{(0)} = 100$, $x_2^{(0)} = 100$, $x_3^{(0)} = 100$, $x_4^{(0)} = 100$, and compute from (2) a perhaps better approximation

Use "old" values
("New" values here not yet available)

$$
(3)\qquad
\begin{aligned}
x_1^{(1)} &= \phantom{0.25x_1^{(1)}} \quad 0.25x_2^{(0)} + 0.25x_3^{(0)} \phantom{{}+ 0.25x_4^{(0)}} + 50.00 = 100.00\\
x_2^{(1)} &= 0.25x_1^{(1)} \phantom{{}+ 0.25x_2^{(1)}} \phantom{{}+ 0.25x_3^{(1)}} \quad 0.25x_4^{(0)} + 50.00 = 100.00\\
x_3^{(1)} &= 0.25x_1^{(1)} \phantom{{}+ 0.25x_2^{(1)}} \phantom{{}+ 0.25x_3^{(1)}} \quad 0.25x_4^{(0)} + 25.00 = 75.00\\
x_4^{(1)} &= \phantom{0.25x_1^{(1)}} \quad 0.25x_2^{(1)} + 0.25x_3^{(1)} \phantom{{}+ 0.25x_4^{(0)}} + 25.00 = 68.75
\end{aligned}
$$

Use "new" values

These equations (3) are obtained from (2) by substituting on the right the ***most recent*** approximation for each unknown. In fact, corresponding values replace previous ones as soon as they have been computed, so that in the second and third equations we use $x_1^{(1)}$ (not $x_1^{(0)}$), and in the last equation of (3) we use $x_2^{(1)}$ and $x_3^{(1)}$ (not $x_2^{(0)}$ and $x_3^{(0)}$). Using the same principle, we obtain in the next step

$$
\begin{aligned}
x_1^{(2)} &= \phantom{0.25x_1^{(2)} +{}} 0.25x_2^{(1)} + 0.25x_3^{(1)} \phantom{{}+ 0.25x_4^{(1)}} + 50.00 = 93.750\\
x_2^{(2)} &= 0.25x_1^{(2)} \phantom{{}+ 0.25x_2^{(2)}} \phantom{{}+ 0.25x_3^{(2)}} + 0.25x_4^{(1)} + 50.00 = 90.625\\
x_3^{(2)} &= 0.25x_1^{(2)} \phantom{{}+ 0.25x_2^{(2)}} \phantom{{}+ 0.25x_3^{(2)}} + 0.25x_4^{(1)} + 25.00 = 65.625\\
x_4^{(2)} &= \phantom{0.25x_1^{(2)} +{}} 0.25x_2^{(2)} + 0.25x_3^{(2)} \phantom{{}+ 0.25x_4^{(1)}} + 25.00 = 64.062
\end{aligned}
$$

Further steps give the values

x_1	x_2	x_3	x_4
89.062	88.281	63.281	62.891
87.891	87.695	62.695	62.598
87.598	87.549	62.549	62.524
87.524	87.512	62.512	62.506
87.506	87.503	62.503	62.502

Hence convergence to the exact solution $x_1 = x_2 = 87.5$, $x_3 = x_4 = 62.5$ (verify!) seems rather fast. ■

An algorithm for the Gauss–Seidel iteration is shown in Table 20.2. To obtain the algorithm, let us derive the general formulas for this iteration.

We assume that $a_{jj} = 1$ for $j = 1, \cdots, n$. (Note that this can be achieved if we can rearrange the equations so that no diagonal coefficient is zero; then we may divide each equation by the corresponding diagonal coefficient.) We now write

www.EngineeringEBooksPdf.com

$$\mathbf{A} = \mathbf{I} + \mathbf{L} + \mathbf{U} \qquad (a_{jj} = 1) \tag{4}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix and $\mathbf{L}$ and $\mathbf{U}$ are, respectively, lower and upper triangular matrices with zero main diagonals. If we substitute (4) into $\mathbf{Ax} = \mathbf{b}$, we have

$$\mathbf{Ax} = (\mathbf{I} + \mathbf{L} + \mathbf{U})\mathbf{x} = \mathbf{b}.$$

Taking $\mathbf{Lx}$ and $\mathbf{Ux}$ to the right, we obtain, since $\mathbf{Ix} = \mathbf{x}$,

$$\mathbf{x} = \mathbf{b} - \mathbf{Lx} - \mathbf{Ux}. \tag{5}$$

Remembering from (3) in Example 1 that below the main diagonal we took "new" approximations and above the main diagonal "old" ones, we obtain from (5) the desired iteration formulas

$$\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{L}\overset{\text{"New"}}{\mathbf{x}^{(m+1)}} - \mathbf{U}\overset{\text{"Old"}}{\mathbf{x}^{(m)}} \qquad (a_{jj} = 1) \tag{6}$$

where $\mathbf{x}^{(m)} = [x_j^{(m)}]$ is the mth approximation and $\mathbf{x}^{(m+1)} = [x_j^{(m+1)}]$ is the $(m + 1)$st approximation. In components this gives the formula in line 1 in Table 20.2. The matrix $\mathbf{A}$ must satisfy $a_{jj} \neq 0$ for all j. In Table 20.2 our assumption $a_{jj} = 1$ is no longer required, but is automatically taken care of by the factor $1/a_{jj}$ in line 1.

Table 20.2 Gauss–Seidel Iteration

ALGORITHM GAUSS–SEIDEL ($\mathbf{A}$, $\mathbf{b}$, $\mathbf{x}^{(0)}$, ϵ, N)

This algorithm computes a solution $\mathbf{x}$ of the system $\mathbf{Ax} = \mathbf{b}$ given an initial approximation $\mathbf{x}^{(0)}$, where $\mathbf{A} = [a_{jk}]$ is an $n \times n$ matrix with $a_{jj} \neq 0$, $j = 1, \cdots, n$.

INPUT: $\mathbf{A}$, $\mathbf{b}$, initial approximation $\mathbf{x}^{(0)}$, tolerance $\epsilon > 0$, maximum number of iterations N

OUTPUT: Approximate solution $\mathbf{x}^{(m)} = [x_j^{(m)}]$ or failure message that $\mathbf{x}^{(N)}$ does not satisfy the tolerance condition

For $m = 0, \cdots, N - 1$, do:

For $j = 1, \cdots, n$, do:

1 $$x_j^{(m+1)} = \frac{1}{a_{jj}}\left(b_j - \sum_{k=1}^{j-1} a_{jk}x_k^{(m+1)} - \sum_{k=j+1}^{n} a_{jk}x_k^{(m)}\right)$$

End

2 If $\max_j |x_j^{(m+1)} - x_j^{(m)}| < \epsilon\, |x_j^{(m+1)}|$ then OUTPUT $\mathbf{x}^{(m+1)}$. Stop

[*Procedure completed successfully*]

End

OUTPUT: "No solution satisfying the tolerance condition obtained after N iteration steps." Stop

[*Procedure completed unsuccessfully*]

End GAUSS–SEIDEL

www.EngineeringEBooksPdf.com

Convergence and Matrix Norms

An iteration method for solving $\mathbf{Ax} = \mathbf{b}$ is said to **converge** for an initial $\mathbf{x}^{(0)}$ if the corresponding iterative sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots$ converges to a solution of the given system. Convergence depends on the relation between $\mathbf{x}^{(m)}$ and $\mathbf{x}^{(m+1)}$. To get this relation for the Gauss–Seidel method, we use (6). We first have

$$(\mathbf{I} + \mathbf{L})\,\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{U}\mathbf{x}^{(m)}$$

and by multiplying by $(\mathbf{I} + \mathbf{L})^{-1}$ from the left,

$$\mathbf{x}^{(m+1)} = \mathbf{C}\mathbf{x}^{(m)} + (\mathbf{I} + \mathbf{L})^{-1}\,\mathbf{b} \qquad \text{where} \qquad \mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\,\mathbf{U}. \tag{7}$$

The Gauss–Seidel iteration converges for every $\mathbf{x}^{(0)}$ if and only if all the eigenvalues (Sec. 8.1) of the "iteration matrix" $\mathbf{C} = [c_{jk}]$ have absolute value less than 1. (Proof in Ref. [E5], p. 191, listed in App. 1.)

CAUTION! If you want to get $\mathbf{C}$, first divide the rows of $\mathbf{A}$ by a_{jj} to have main diagonal $1, \cdots, 1$. If the **spectral radius** of $\mathbf{C}$ (= maximum of those absolute values) is small, then the convergence is rapid.

Sufficient Convergence Condition. A sufficient condition for convergence is

$$\|\mathbf{C}\| < 1. \tag{8}$$

Here $\|\mathbf{C}\|$ is some **matrix norm**, such as

$$\|\mathbf{C}\| = \sqrt{\sum_{j=1}^{n} \sum_{k=1}^{n} c_{jk}^2} \qquad \textbf{(Frobenius norm)} \tag{9}$$

or the greatest of the sums of the $|c_{jk}|$ in a *column* of $\mathbf{C}$

$$\|\mathbf{C}\| = \max_k \sum_{j=1}^{n} |c_{jk}| \qquad \textbf{(Column "sum" norm)} \tag{10}$$

or the greatest of the sums of the $|c_{jk}|$ in a *row* of $\mathbf{C}$

$$\|\mathbf{C}\| = \max_j \sum_{k=1}^{n} |c_{jk}| \qquad \textbf{(Row "sum" norm)}. \tag{11}$$

These are the most frequently used matrix norms in numerics.

In most cases the choice of one of these norms is a matter of computational convenience. However, the following example shows that sometimes one of these norms is preferable to the others.

www.EngineeringEBooksPdf.com

EXAMPLE 2 Test of Convergence of the Gauss–Seidel Iteration

Test whether the Gauss–Seidel iteration converges for the system

$$\begin{aligned} 2x + y + z &= 4 \\ x + 2y + z &= 4 \\ x + y + 2z &= 4 \end{aligned} \qquad \text{written} \qquad \begin{aligned} x &= 2 - \tfrac{1}{2}y - \tfrac{1}{2}z \\ y &= 2 - \tfrac{1}{2}x - \tfrac{1}{2}z \\ z &= 2 - \tfrac{1}{2}x - \tfrac{1}{2}y. \end{aligned}$$

Solution. The decomposition (multiply the matrix by $\frac{1}{2}$ – why?) is

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} = \mathbf{I} + \mathbf{L} + \mathbf{U} = \mathbf{I} + \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} + \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix}.$$

It shows that

$$\mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U} = -\begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{4} & -\frac{1}{2} & 1 \end{bmatrix}\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{4} & -\frac{1}{4} \\ 0 & \frac{1}{8} & \frac{3}{8} \end{bmatrix}.$$

We compute the Frobenius norm of $\mathbf{C}$

$$\|\mathbf{C}\| = \left(\tfrac{1}{4} + \tfrac{1}{4} + \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{64} + \tfrac{9}{64}\right)^{1/2} = \left(\tfrac{50}{64}\right)^{1/2} = 0.884 < 1$$

and conclude from (8) that this Gauss–Seidel iteration converges. It is interesting that the other two norms would permit no conclusion, as you should verify. Of course, this points to the fact that (8) is sufficient for convergence rather than necessary. ■

Residual. Given a system $\mathbf{Ax} = \mathbf{b}$, the **residual** $\mathbf{r}$ of $\mathbf{x}$ with respect to this system is defined by

$$\mathbf{r} = \mathbf{b} - \mathbf{Ax}. \tag{12}$$

Clearly, $\mathbf{r} = \mathbf{0}$ if and only if $\mathbf{x}$ is a solution. Hence $\mathbf{r} \neq \mathbf{0}$ for an approximate solution. In the Gauss–Seidel iteration, at each stage we modify or *relax* a component of an approximate solution in order to reduce a component of $\mathbf{r}$ to zero. Hence the Gauss–Seidel iteration belongs to a class of methods often called **relaxation methods**. More about the residual follows in the next section.

Jacobi Iteration

The Gauss–Seidel iteration is a method of **successive corrections** because for each component we successively replace an approximation of a component by a corresponding new approximation as soon as the latter has been computed. An iteration method is called a method of **simultaneous corrections** if no component of an approximation $\mathbf{x}^{(m)}$ is used until ***all*** the components of $\mathbf{x}^{(m)}$ have been computed. A method of this type is the **Jacobi iteration**, which is similar to the Gauss–Seidel iteration but involves *not* using improved values until a step has been completed and then replacing $\mathbf{x}^{(m)}$ by $\mathbf{x}^{(m+1)}$ at once, directly before the beginning of the next step. Hence if we write $\mathbf{Ax} = \mathbf{b}$ (***with*** $a_{jj} = 1$ ***as before!***) in the form $\mathbf{x} = \mathbf{b} + (\mathbf{I} - \mathbf{A})\mathbf{x}$, the Jacobi iteration in matrix notation is

$$\mathbf{x}^{(m+1)} = \mathbf{b} + (\mathbf{I} - \mathbf{A})\mathbf{x}^{(m)} \qquad (a_{jj} = 1). \tag{13}$$

www.EngineeringEBooksPdf.com

This method converges for every choice of $\mathbf{x}^{(0)}$ if and only if the spectral radius of $\mathbf{I} - \mathbf{A}$ is less than 1. It has recently gained greater practical interest since on parallel processors all n equations can be solved simultaneously at each iteration step.

For Jacobi, see Sec. 10.3. For exercises, see the problem set.

PROBLEM SET 20.3

1. Verify the solution in Example 1 of the text.

2. Show that for the system in Example 2 the Jacobi iteration diverges. *Hint.* Use eigenvalues.

3. Verify the claim at the end of Example 2.

4–10 GAUSS–SEIDEL ITERATION

Do 5 steps, starting from $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$ and using 6S in the computation. *Hint.* Make sure that you solve each equation for the variable that has the largest coefficient (why?). Show the details.

4.
$$\begin{aligned} 4x_1 - x_2 \phantom{{}- x_3} &= 21 \\ -x_1 + 4x_2 - x_3 &= -45 \\ - x_2 + 4x_3 &= 33 \end{aligned}$$

5.
$$\begin{aligned} 10x_1 + x_2 + x_3 &= 6 \\ x_1 + 10x_2 + x_3 &= 6 \\ x_1 + x_2 + 10x_3 &= 6 \end{aligned}$$

6.
$$\begin{aligned} x_2 + 7x_3 &= 25.5 \\ 5x_1 + x_2 \phantom{{}+ 7x_3} &= 0 \\ x_1 + 6x_2 + x_3 &= -10.5 \end{aligned}$$

7.
$$\begin{aligned} 5x_1 - 2x_2 \phantom{{}- 2x_3} &= 18 \\ -2x_1 + 10x_2 - 2x_3 &= -60 \\ - 2x_2 + 15x_3 &= 128 \end{aligned}$$

8.
$$\begin{aligned} 3x_1 + 2x_2 + x_3 &= 7 \\ x_1 + 3x_2 + 2x_3 &= 4 \\ 2x_1 + x_2 + 3x_3 &= 7 \end{aligned}$$

9.
$$\begin{aligned} 5x_1 + x_2 + 2x_3 &= 19 \\ x_1 + 4x_2 - 2x_3 &= -2 \\ 2x_1 + 3x_2 + 8x_3 &= 39 \end{aligned}$$

10.
$$\begin{aligned} 4x_1 \phantom{{}+ 6x_2} + 5x_3 &= 12.5 \\ x_1 + 6x_2 + 2x_3 &= 18.5 \\ 8x_1 + 2x_2 + x_3 &= -11.5 \end{aligned}$$

11. Apply the Gauss–Seidel iteration (3 steps) to the system in Prob. 5, starting from **(a)** 0, 0, 0 **(b)** 10, 10, 10. Compare and comment.

12. In Prob. 5, compute **C** **(a)** if you solve the first equation for x_1, the second for x_2, the third for x_3, proving convergence; **(b)** if you nonsensically solve the third equation for x_1, the first for x_2, the second for x_3, proving divergence.

13. CAS Experiment. Gauss–Seidel Iteration. (a) Write a program for Gauss–Seidel iteration.

(b) Apply the program $\mathbf{A}(t)\mathbf{x} = \mathbf{b}$, to starting from $[0 \quad 0 \quad 0]^\mathsf{T}$, where

$$\mathbf{A}(t) = \begin{bmatrix} 1 & t & t \\ t & 1 & t \\ t & t & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}.$$

For $t = 0.2, 0.5, 0.8, 0.9$ determine the number of steps to obtain the exact solution to 6S and the corresponding spectral radius of **C**. Graph the number of steps and the spectral radius as functions of t and comment.

(c) Successive overrelaxation (SOR). Show that by adding and subtracting $\mathbf{x}^{(m)}$ on the right, formula (6) can be written

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - (\mathbf{U} + \mathbf{I})\mathbf{x}^{(m)} \qquad (a_{jj} = 1).$$

Anticipation of further corrections motivates the introduction of an **overrelaxation factor** $\omega > 1$ to get the **SOR formula for Gauss–Seidel**

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \omega(\mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - (\mathbf{U} + \mathbf{I})\mathbf{x}^{(m)}) \qquad (a_{jj} = 1) \tag{14}$$

intended to give more rapid convergence. A recommended value is $\omega = 2/(1 + \sqrt{1 - \rho})$, where ρ is the spectral radius of **C** in (7). Apply SOR to the matrix in (b) for $t = 0.5$ and 0.8 and notice the improvement of convergence. (Spectacular gains are made with larger systems.)

www.EngineeringEBooksPdf.com

14–17 JACOBI ITERATION

Do 5 steps, starting from $\mathbf{x}_0 = [1 \quad 1 \quad 1]$. Compare with the Gauss–Seidel iteration. Which of the two seems to converge faster? Show the details of your work.

14. The system in Prob. 4

15. The system in Prob. 9

16. The system in Prob. 10

17. Show convergence in Prob. 16 by verifying that $\mathbf{I} - \mathbf{A}$, where $\mathbf{A}$ is the matrix in Prob. 16 with the rows divided by the corresponding main diagonal entries, has the eigenvalues -0.519589 and $0.259795 \pm 0.246603i$.

18–20 NORMS

Compute the norms (9), (10), (11) for the following (square) matrices. Comment on the reasons for greater or smaller differences among the three numbers.

18. The matrix in Prob. 10

19. The matrix in Prob. 5

20. $\begin{bmatrix} 2k & -k & -k \\ k & -2k & k \\ -k & -k & 2k \end{bmatrix}$

20.4 Linear Systems: Ill-Conditioning, Norms

One does not need much experience to observe that some systems $\mathbf{Ax} = \mathbf{b}$ are good, giving accurate solutions even under roundoff or coefficient inaccuracies, whereas others are bad, so that these inaccuracies affect the solution strongly. We want to see what is going on and whether or not we can "trust" a linear system. Let us first formulate the two relevant concepts (ill- and well-conditioned) for general numeric work and then turn to linear systems and matrices.

A computational problem is called **ill-conditioned** (or *ill-posed*) if "small" changes in the data (the input) cause "large" changes in the solution (the output). On the other hand, a problem is called **well-conditioned** (or *well-posed*) if "small" changes in the data cause only "small" changes in the solution.

These concepts are qualitative. We would certainly regard a magnification of inaccuracies by a factor 100 as "large," but could debate where to draw the line between "large" and "small," depending on the kind of problem and on our viewpoint. Double precision may sometimes help, but if data are measured inaccurately, one should attempt ***changing the mathematical setting*** of the problem to a well-conditioned one.

Let us now turn to linear systems. Figure 445 explains that ill-conditioning occurs if and only if the two equations give two nearly parallel lines, so that their intersection point (the solution of the system) moves substantially if we raise or lower a line just a little. For larger systems the situation is similar in principle, although geometry no longer helps. We shall see that we may regard ill-conditioning as an approach to singularity of the matrix.

Fig. 445. (a) Well-conditioned and (b) ill-conditioned linear system of two equations in two unknowns

www.EngineeringEBooksPdf.com

EXAMPLE 1 An Ill-Conditioned System

You may verify that the system

$$\begin{aligned} 0.9999x - 1.0001y &= 1 \\ x - \qquad y &= 1 \end{aligned}$$

has the solution $x = 0.5$, $y = -0.5$, whereas the system

$$\begin{aligned} 0.9999x - 1.0001y &= 1 \\ x - \qquad y &= 1 + \epsilon \end{aligned}$$

has the solution $x = 0.5 + 5000.5\epsilon$, $y = -0.5 + 4999.5\epsilon$. This shows that the system is ill-conditioned because a change on the right of magnitude ϵ produces a change in the solution of magnitude 5000ϵ, approximately. We see that the lines given by the equations have nearly the same slope. ■

Well-conditioning can be asserted if the main diagonal entries of $\mathbf{A}$ have large absolute values compared to those of the other entries. Similarly if $\mathbf{A}^{-1}$ and $\mathbf{A}$ have maximum entries of about the same absolute value.

Ill-conditioning is indicated if $\mathbf{A}^{-1}$ has entries of large absolute value compared to those of the solution (about 5000 in Example 1) and if poor approximate solutions may still produce small residuals.

Residual. The *residual* $\mathbf{r}$ of an approximate solution $\tilde{\mathbf{x}}$ of $\mathbf{Ax} = \mathbf{b}$ is defined as

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}. \tag{1}$$

Now $\mathbf{b} = \mathbf{Ax}$, so that

$$\mathbf{r} = \mathbf{A}(\mathbf{x} - \mathbf{A}\tilde{\mathbf{x}}). \tag{2}$$

Hence $\mathbf{r}$ is small if $\tilde{\mathbf{x}}$ has high accuracy, but the converse may be false:

EXAMPLE 2 Inaccurate Approximate Solution with a Small Residual

The system

$$\begin{aligned} 1.0001x_1 + \qquad x_2 &= 2.0001 \\ x_1 + 1.0001x_2 &= 2.0001 \end{aligned}$$

has the exact solution $x_1 = 1, x_2 = 1$. Can you see this by inspection? The very inaccurate approximation $\tilde{x}_1 = 2.0000, \tilde{x}_2 = 0.0001$ has the very small residual (to 4D)

$$\mathbf{r} = \begin{bmatrix} 2.0001 \\ 2.0001 \end{bmatrix} - \begin{bmatrix} 1.0001 & 1.0000 \\ 1.0000 & 1.0001 \end{bmatrix} \begin{bmatrix} 2.0000 \\ 0.0001 \end{bmatrix} = \begin{bmatrix} 2.0001 \\ 2.0001 \end{bmatrix} - \begin{bmatrix} 2.0003 \\ 2.0001 \end{bmatrix} = \begin{bmatrix} -0.0002 \\ 0.0000 \end{bmatrix}.$$

From this, a naive person might draw the false conclusion that the approximation should be accurate to 3 or 4 decimals.

Our result is probably unexpected, but we shall see that it has to do with the fact that the system is ill-conditioned. ■

Our goal is to show that ill-conditioning of a linear system and of its coefficient matrix $\mathbf{A}$ can be measured by a number, the *condition number* $\kappa(\mathbf{A})$. Other measures for ill-conditioning

www.EngineeringEBooksPdf.com

have also been proposed, but $\kappa(\mathbf{A})$ is probably the most widely used one. $\kappa(\mathbf{A})$ is defined in terms of norm, a concept of great general interest throughout numerics (and in modern mathematics in general!). We shall reach our goal in three steps, discussing

1. **Vector norms**
2. **Matrix norms**
3. **Condition number** κ of a square matrix

Vector Norms

A **vector norm** for column vectors $\mathbf{x} = [x_j]$ with n components (n fixed) is a generalized length or distance. It is denoted by $\|\mathbf{x}\|$ and is defined by four properties of the usual length of vectors in three-dimensional space, namely,

$$
(3)\qquad
\begin{aligned}
&\text{(a)}\quad \|\mathbf{x}\| \text{ is a nonnegative real number.}\\
&\text{(b)}\quad \|\mathbf{x}\| = 0 \quad \text{if and only if} \quad \mathbf{x} = \mathbf{0}.\\
&\text{(c)}\quad \|k\mathbf{x}\| = |k|\,\|\mathbf{x}\| \qquad \text{for all } k.\\
&\text{(d)}\quad \|\mathbf{x} + \mathbf{y}\| \leqq \|\mathbf{x}\| + \|\mathbf{y}\| \qquad \text{(Triangle inequality).}
\end{aligned}
$$

If we use several norms, we label them by a subscript. Most important in connection with computations is the ***p-norm*** defined by

$$
(4)\qquad \|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}
$$

where p is a fixed number and $p \geqq 1$. In practice, one usually takes $p = 1$ or 2 and, as a third norm, $\|\mathbf{x}\|_\infty$ (the latter as defined below), that is,

$$
\boxed{
\begin{aligned}
&\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n| && (\text{“}l_1\text{-norm”})\\
&\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} && (\text{“Euclidean” or “}l_2\text{-norm”})\\
&\|\mathbf{x}\|_\infty = \max_j |x_j| && (\text{“}l_\infty\text{-norm”}).
\end{aligned}
}
\qquad \begin{aligned}&\textbf{(5)}\\ &\textbf{(6)}\\ &\textbf{(7)}\end{aligned}
$$

For $n = 3$ the l_2-norm is the usual length of a vector in three-dimensional space. The l_1-norm and l_∞-norm are generally more convenient in computation. But all three norms are in common use.

EXAMPLE 3 Vector Norms

If $\mathbf{x}^\mathsf{T} = [2 \quad -3 \quad 0 \quad 1 \quad -4]$, then $\|\mathbf{x}\|_1 = 10$, $\|\mathbf{x}\|_2 = \sqrt{30}$, $\|\mathbf{x}\|_\infty = 4$. ■

In three-dimensional space, two points with position vectors $\mathbf{x}$ and $\tilde{\mathbf{x}}$ have distance $|\mathbf{x} - \tilde{\mathbf{x}}|$ from each other. For a linear system $\mathbf{Ax} = \mathbf{b}$, this suggests that we take $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ as a measure of inaccuracy and call it the **distance** between an exact and an approximate solution, or the **error** of $\tilde{\mathbf{x}}$.

Matrix Norm

If $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{x}$ any vector with n components, then $\mathbf{Ax}$ is a vector with n components. We now take a vector norm and consider $\|\mathbf{x}\|$ and $\|\mathbf{Ax}\|$. One can prove (see

www.EngineeringEBooksPdf.com

Ref. [E17]. pp. 77, 92–93, listed in App. 1) that there is a number c (depending on $\mathbf{A}$) such that

$$\|\mathbf{Ax}\| \leqq c\|\mathbf{x}\| \qquad \text{for all } \mathbf{x}. \tag{8}$$

Let $\mathbf{x} \neq 0$. Then $\|\mathbf{x}\| > 0$ by (3b) and division gives $\|\mathbf{Ax}\|/\|\mathbf{x}\| \leqq c$. We obtain the smallest possible c valid for *all* $\mathbf{x}$ ($\neq \mathbf{0}$) by taking the maximum on the left. This smallest c is called the **matrix norm of A** *corresponding to the vector norm we picked* and is denoted by $\|\mathbf{A}\|$. Thus

$$\|\mathbf{A}\| = \max \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \qquad (\mathbf{x} \neq \mathbf{0}), \tag{9}$$

the maximum being taken over all $\mathbf{x} \neq \mathbf{0}$. Alternatively [see (c) in Team Project 24],

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|. \tag{10}$$

The maximum in (10) and thus also in (9) exists. And the name "matrix ***norm***" is justified because $\|\mathbf{A}\|$ satisfies (3) with $\mathbf{x}$ and $\mathbf{y}$ replaced by $\mathbf{A}$ and $\mathbf{B}$. (Proofs in Ref. [E17] pp. 77, 92–93.)

Note carefully that $\|\mathbf{A}\|$ depends on the vector norm that we selected. In particular, one can show that

for the l_1-norm (5) one gets the column "sum" norm (10), Sec. 20.3,

for the l_∞-norm (7) one gets the row "sum" norm (11), Sec. 20.3.

By taking our best possible (our smallest) $c = \|\mathbf{A}\|$ we have from (8)

$$\|\mathbf{Ax}\| \leqq \|\mathbf{A}\|\,\|\mathbf{x}\|. \tag{11}$$

This is the formula we shall need. Formula (9) also implies for two $n \times n$ matrices (see Ref. [E17], p. 98)

$$\|\mathbf{AB}\| \leqq \|\mathbf{A}\|\,\|\mathbf{B}\|, \qquad \text{thus} \qquad \|\mathbf{A}^n\| \leqq \|\mathbf{A}\|^n. \tag{12}$$

See Refs. [E9] and [E17] for other useful formulas on norms.

Before we go on, let us do a simple illustrative computation.

EXAMPLE 4 Matrix Norms

Compute the matrix norms of the coefficient matrix $\mathbf{A}$ in Example 1 and of its inverse $\mathbf{A}^{-1}$, assuming that we use (a) the l_1-vector norm, (b) the l_∞-vector norm.

Solution. We use (4*), Sec. 7.8, for the inverse and then (10) and (11) in Sec. 20.3. Thus

$$\mathbf{A} = \begin{bmatrix} 0.9999 & -1.0001 \\ 1.0000 & -1.0000 \end{bmatrix}, \qquad \mathbf{A}^{-1} = \begin{bmatrix} -5000.0 & 5000.5 \\ -5000.0 & 4999.5 \end{bmatrix}.$$

(a) The l_1-vector norm gives the column "sum" norm (10), Sec. 20.3; from Column 2 we thus obtain $\|\mathbf{A}\| = |-1.0001| + |-1.0000| = 2.0001$. Similarly, $\|\mathbf{A}^{-1}\| = 10{,}000$.

www.EngineeringEBooksPdf.com

(b) The l_∞-vector norm gives the row "sum" norm (11), Sec. 20.3; thus $\|\mathbf{A}\| = 2$, $\|\mathbf{A}^{-1}\| = 10000.5$ from Row 1. We notice that $\|\mathbf{A}^{-1}\|$ is surprisingly large, which makes the product $\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ large (20,001). We shall see below that this is typical of an ill-conditioned system.

Condition Number of a Matrix

We are now ready to introduce the key concept in our discussion of ill-conditioning, the **condition number** $\kappa(\mathbf{A})$ of a (nonsingular) square matrix $\mathbf{A}$, defined by

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|. \tag{13}$$

The role of the condition number is seen from the following theorem.

THEOREM 1 **Condition Number**

A linear system of equations $\mathbf{Ax} = \mathbf{b}$ *and its matrix* $\mathbf{A}$ *whose condition number* (13) *is small are well-conditioned. A large condition number indicates ill-conditioning.*

PROOF $\mathbf{b} = \mathbf{Ax}$ and (11) give $\|\mathbf{b}\| \leqq \|\mathbf{A}\|\,\|\mathbf{x}\|$. Let $\mathbf{b} \neq \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$. Then division by $\|\mathbf{b}\|\,\|\mathbf{x}\|$ gives

$$\frac{1}{\|\mathbf{x}\|} \leqq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}. \tag{14}$$

Multiplying (2) $\mathbf{r} = \mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}})$ by $\mathbf{A}^{-1}$ from the left and interchanging sides, we have $\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{r}$. Now (11) with $\mathbf{A}^{-1}$ and $\mathbf{r}$ instead of $\mathbf{A}$ and $\mathbf{x}$ yields

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|\mathbf{A}^{-1}\mathbf{r}\| \leqq \|\mathbf{A}^{-1}\|\,\|\mathbf{r}\|.$$

Division by $\|\mathbf{x}\|$ [note that $\|\mathbf{x}\| \neq 0$ by (3b)] and use of (14) finally gives

$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leqq \frac{1}{\|\mathbf{x}\|}\,\|\mathbf{A}^{-1}\|\,\|\mathbf{r}\| \leqq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}\,\|\mathbf{A}^{-1}\|\,\|\mathbf{r}\| = \kappa(\mathbf{A})\,\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}. \tag{15}$$

Hence if $\kappa(\mathbf{A})$ is small, a small $\|\mathbf{r}\|/\|\mathbf{b}\|$ implies a small relative error $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$, so that the system is well-conditioned. However, this does not hold if $\kappa(\mathbf{A})$ is large; then a small $\|\mathbf{r}\|/\|\mathbf{b}\|$ does not necessarily imply a small relative error $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$. ■

EXAMPLE 5 **Condition Numbers. Gauss–Seidel Iteration**

$$\mathbf{A} = \begin{bmatrix} 5 & 1 & 1 \\ 1 & 4 & 2 \\ 1 & 2 & 4 \end{bmatrix} \quad \text{has the inverse} \quad \mathbf{A}^{-1} = \frac{1}{56}\begin{bmatrix} 12 & -2 & -2 \\ -2 & 19 & -9 \\ -2 & -9 & 19 \end{bmatrix}.$$

Since $\mathbf{A}$ is symmetric, (10) and (11) in Sec. 20.3 give the same condition number

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\| = 7 \cdot \tfrac{1}{56} \cdot 30 = 3.75.$$

We see that a linear system $\mathbf{Ax} = \mathbf{b}$ with this $\mathbf{A}$ is well-conditioned.

www.EngineeringEBooksPdf.com

For instance, if $\mathbf{b} = [14 \quad 0 \quad 28]^T$, the Gauss algorithm gives the solution $\mathbf{x} = [2 \quad -5 \quad 9]^T$, (confirm this). Since the main diagonal entries of $\mathbf{A}$ are relatively large, we can expect reasonably good convergence of the Gauss–Seidel iteration. Indeed, starting from, say, $\mathbf{x}_0 = [1 \quad 1 \quad 1]^T$, we obtain the first 8 steps (3D values)

x_1	x_2	x_3
1.000	1.000	1.000
2.400	−1.100	6.950
1.630	−3.882	8.534
1.870	−4.734	8.900
1.967	−4.942	8.979
1.993	−4.988	8.996
1.998	−4.997	8.999
2.000	−5.000	9.000
2.000	−5.000	9.000

■

EXAMPLE 6 Ill-Conditioned Linear System

Example 4 gives by (10) or (11), Sec. 20.3, for the matrix in Example 1 the very large condition number $\kappa(\mathbf{A}) = 2.0001 \cdot 10000 = 2 \cdot 10000.5 = 200001$. This confirms that the system is very ill-conditioned.

Similarly in Example 2, where by (4*), Sec. 7.8 and 6D-computation,

$$\mathbf{A}^{-1} = \frac{1}{0.0002}\begin{bmatrix} 1.0001 & -1.0000 \\ -1.0000 & 1.0001 \end{bmatrix} = \begin{bmatrix} 5000.5 & -5.000.0 \\ -5000.0 & 5000.5 \end{bmatrix}$$

so that (10), Sec. 20.3, gives a very large $\kappa(\mathbf{A})$, explaining the surprising result in Example 2,

$$\kappa(\mathbf{A}) = (1.0001 + 1.0000)(5000.5 + 5000.0) \approx 20{,}002.$$

■

In practice, $\mathbf{A}^{-1}$ will not be known, so that in computing the condition number $\kappa(\mathbf{A})$, one must estimate $\|\mathbf{A}^{-1}\|$. A method for this (proposed in 1979) is explained in Ref. [E9] listed in App. 1.

Inaccurate Matrix Entries. $\kappa(\mathbf{A})$ can be used for estimating the effect $\delta\mathbf{x}$ of an inaccuracy $\delta\mathbf{A}$ of $\mathbf{A}$ (errors of measurements of the a_{jk}, for instance). Instead of $\mathbf{Ax} = \mathbf{b}$ we then have

$$(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}.$$

Multiplying out and subtracting $\mathbf{Ax} = \mathbf{b}$ on both sides, we obtain

$$\mathbf{A}\delta\mathbf{x} + \delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{0}.$$

Multiplication by $\mathbf{A}^{-1}$ from the left and taking the second term to the right gives

$$\delta\mathbf{x} = -\mathbf{A}^{-1}\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}).$$

Applying (11) with $\mathbf{A}^{-1}$ and vector $\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})$ instead of $\mathbf{A}$ and $\mathbf{x}$, we get

$$\|\delta\mathbf{x}\| = \|\mathbf{A}^{-1}\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})\| \leqq \|\mathbf{A}^{-1}\|\,\|\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})\|.$$

Applying (11) on the right, with $\delta\mathbf{A}$ and $\mathbf{x} - \delta\mathbf{x}$ instead of $\mathbf{A}$ and $\mathbf{x}$, we obtain

$$\|\delta\mathbf{x}\| \leqq \|\mathbf{A}^{-1}\|\,\|\delta\mathbf{A}\|\,\|\mathbf{x} + \delta\mathbf{x}\|.$$

www.EngineeringEBooksPdf.com

Now $\|\mathbf{A}^{-1}\| = \kappa(\mathbf{A})/\|\mathbf{A}\|$ by the definition of $\kappa(\mathbf{A})$, so that division by $\|\mathbf{x} + \delta\mathbf{x}\|$ shows that the relative inaccuracy of $\mathbf{x}$ is related to that of $\mathbf{A}$ via the condition number by the inequality

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \approx \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x} + \delta\mathbf{x}\|} \leqq \|\mathbf{A}^{-1}\|\,\|\delta\mathbf{A}\| = \kappa(\mathbf{A})\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}. \tag{16}$$

Conclusion. If the system is well-conditioned, small inaccuracies $\|\delta\mathbf{A}\|/\|\mathbf{A}\|$ can have only a small effect on the solution. However, in the case of ill-conditioning, if $\|\delta\mathbf{A}\|/\|\mathbf{A}\|$ is small, $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ ***may*** be large.

Inaccurate Right Side. You may show that, similarly, when $\mathbf{A}$ is accurate, an inaccuracy $\delta\mathbf{b}$ of $\mathbf{b}$ causes an inaccuracy $\delta\mathbf{x}$ satisfying

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq \kappa(\mathbf{A})\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{17}$$

Hence $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ must remain relatively small whenever $\kappa(\mathbf{A})$ is small.

EXAMPLE 7 Inaccuracies. Bounds (16) and (17)

If each of the nine entries of $\mathbf{A}$ in Example 5 is measured with an inaccuracy of 0.1, then $\|\delta\mathbf{A}\| = 9 \cdot 0.1$ and (16) gives

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq 7.5 \cdot \frac{3 \cdot 0.1}{7} = 0.321 \qquad \text{thus} \qquad \|\delta\mathbf{x}\| \leqq 0.321\,\|\mathbf{x}\| = 0.321 \cdot 16 = 5.14.$$

By experimentation you will find that the actual inaccuracy $\|\delta\mathbf{x}\|$ is only about 30% of the bound 5.14. This is typical.

Similarly, if $\delta\mathbf{b} = [0.1 \quad 0.1 \quad 0.1]^{\mathsf{T}}$, then $\|\delta\mathbf{b}\| = 0.3$ and $\|\mathbf{b}\| = 42$ in Example 5, so that (17) gives

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq 7.5 \cdot \frac{0.3}{42} = 0.0536, \qquad \text{hence} \qquad \|\delta\mathbf{x}\| \leqq 0.0536 \cdot 16 = 0.857$$

but this bound is again much greater than the actual inaccuracy, which is about 0.15. ■

Further Comments on Condition Numbers. The following additional explanations may be helpful.

1. There is no sharp dividing line between "well-conditioned" and "ill-conditioned," but generally the situation will get worse as we go from systems with small $\kappa(\mathbf{A})$ to systems with larger $\kappa(\mathbf{A})$. Now always $\kappa(\mathbf{A}) \geqq 1$, so that values of 10 or 20 or so give no reason for concern, whereas $\kappa(\mathbf{A}) = 100$, say, calls for caution, and systems such as those in Examples 1 and 2 are extremely ill-conditioned.

2. If $\kappa(\mathbf{A})$ is large (or small) in one norm, it will be large (or small, respectively) in any other norm. See Example 5.

3. The literature on ill-conditioning is extensive. For an introduction to it, see [E9].

This is the end of our discussion of numerics for solving linear systems. In the next section we consider curve fitting, an important area in which solutions are obtained from linear systems.

www.EngineeringEBooksPdf.com

PROBLEM SET 20.4

1–6 VECTOR NORMS

Compute the norms (5), (6), (7). Compute a corresponding **unit vector** (vector of norm 1) with respect to the l_∞-norm.

1. $[1 \quad -3 \quad 8 \quad 0 \quad -6 \quad 0]$

2. $[4 \quad -1 \quad 8]$

3. $[0.2 \quad 0.6 \quad -2.1 \quad 3.0]$

4. $[k^2, \quad 4k, \quad k^3], \quad k > 4$

5. $[1 \quad 1 \quad 1 \quad 1 \quad 1]$

6. $[0 \quad 0 \quad 0 \quad 1 \quad 0]$

7. For what $\mathbf{x} = [a \quad b \quad c]$ will $\|\mathbf{x}\|_1 = \|\mathbf{x}\|_2$?

8. Show that $\|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1$.

9–16 MATRIX NORMS, CONDITION NUMBERS

Compute the matrix norm and the condition number corresponding to the l_1-vector norm.

9. $\begin{bmatrix} 2 & 1 \\ 0 & 4 \end{bmatrix}$

10. $\begin{bmatrix} 2.1 & 4.5 \\ 0.5 & 1.8 \end{bmatrix}$

11. $\begin{bmatrix} \sqrt{5} & 5 \\ 0 & -\sqrt{5} \end{bmatrix}$

12. $\begin{bmatrix} 7 & 6 \\ 6 & 5 \end{bmatrix}$

13. $\begin{bmatrix} -2 & 4 & -1 \\ -2 & 3 & 0 \\ 7 & -12 & 2 \end{bmatrix}$

14. $\begin{bmatrix} 1 & 0.01 & 0 \\ 0.01 & 1 & 0.01 \\ 0 & 0.01 & 1 \end{bmatrix}$

15. $\begin{bmatrix} -20 & 0 & 0 \\ 0 & 0.05 & 0 \\ 0 & 0 & 20 \end{bmatrix}$

16. $\begin{bmatrix} 21 & 10.5 & 7 & 5.25 \\ 10.5 & 7 & 5.25 & 4.2 \\ 7 & 5.25 & 4.2 & 3.5 \\ 5.25 & 4.2 & 3.5 & 3 \end{bmatrix}$

17. Verify (11) for $\mathbf{x} = [3 \quad 15 \quad -4]^\mathsf{T}$ taken with the l_∞-norm and the matrix in Prob. 13.

18. Verify (12) for the matrices in Probs. 9 and 10.

19–20 ILL-CONDITIONED SYSTEMS

Solve $\mathbf{Ax} = \mathbf{b}_1$, $\mathbf{Ax} = \mathbf{b}_2$. Compare the solutions and comment. Compute the condition number of $\mathbf{A}$.

19. $\mathbf{A} = \begin{bmatrix} 4.50 & 3.55 \\ 3.55 & 2.80 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{bmatrix} 5.2 \\ 4.1 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 5.2 \\ 4.0 \end{bmatrix}$

20. $\mathbf{A} = \begin{bmatrix} 3.0 & 1.7 \\ 1.7 & 1.0 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{bmatrix} 4.7 \\ 2.7 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 4.7 \\ 2.71 \end{bmatrix}$

21. Residual. For $\mathbf{Ax} = \mathbf{b}_1$ in Prob. 19 guess what the residual of $\tilde{\mathbf{x}} = [-10.0 \quad 14.1]^\mathsf{T}$, very poorly approximating $[-2 \quad 4]^\mathsf{T}$, might be. Then calculate and comment.

22. Show that $\kappa(\mathbf{A}) \geqq 1$ for the matrix norms (10), (11), Sec. 20.3, and $\kappa(\mathbf{A}) \geqq \sqrt{n}$ for the Frobenius norm (9), Sec. 20.3.

23. CAS EXPERIMENT. Hilbert Matrices. The 3×3 Hilbert matrix is

$$\mathbf{H}_3 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}.$$

The $n \times n$ Hilbert matrix is $\mathbf{H}_n = [h_{jk}]$, where $h_{jk} = 1/(j + k - 1)$. (Similar matrices occur in curve fitting by least squares.) Compute the condition number $\kappa(\mathbf{H}_n)$ for the matrix norm corresponding to the l_∞- (or l_1-) vector norm, for $n = 2, 3, \cdots, 6$ (or further if you wish). Try to find a formula that gives reasonable approximate values of these rapidly growing numbers.

Solve a few linear systems of your choice, involving an $\mathbf{H}_n$.

24. TEAM PROJECT. Norms. (a) Vector norms in our text are **equivalent**, that is, they are related by double inequalities; for instance,

$$\text{(18)} \qquad \begin{aligned} &\text{(a)} \quad \|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_1 \leqq n\|\mathbf{x}\|_\infty \\ &\text{(b)} \quad \frac{1}{n}\|\mathbf{x}\|_1 \leqq \|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_1. \end{aligned}$$

Hence if for some $\mathbf{x}$, one norm is large (or small), the other norm must also be large (or small). Thus in many investigations the particular choice of a norm is not essential. Prove (18).

(b) The Cauchy–Schwarz inequality is

$$|\mathbf{x}^\mathsf{T}\mathbf{y}| \leqq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

www.EngineeringEBooksPdf.com

It is very important. (Proof in Ref. [GenRef7] listed in App. 1.) Use it to prove

(19a) $$\|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1 \leqq \sqrt{n}\,\|\mathbf{x}\|_2$$

(19b) $$\frac{1}{\sqrt{n}}\|\mathbf{x}\|_1 \leqq \|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1.$$

(c) Formula (10) is often more practical than (9). Derive (10) from (9).

(d) Matrix norms. Illustrate (11) with examples. Give examples of (12) with equality as well as with strict inequality. Prove that the matrix norms (10), (11) in Sec. 20.3 satisfy the *axioms of a norm*

$$\|\mathbf{A}\| \geqq \mathbf{0}.$$
$$\|\mathbf{A}\| = \mathbf{0} \text{ if and only if } \mathbf{A} = \mathbf{0},$$
$$\|k\mathbf{A}\| = |k|\,\|\mathbf{A}\|,$$
$$\|\mathbf{A} + \mathbf{B}\| \leqq \|\mathbf{A}\| + \|\mathbf{B}\|.$$

25. WRITING PROJECT. Norms and Their Use in This Section. Make a list of the most important of the many ideas covered in this section and write a two-page report on them.

20.5 Least Squares Method

Having discussed numerics for linear systems, we now turn to an important application, curve fitting, in which the solutions are obtained from linear systems.

In **curve fitting** we are given n points (pairs of numbers) $(x_1, y_1), \cdots, (x_n, y_n)$ and we want to determine a function $f(x)$ such that

$$f(x_1) \approx y_1, \cdots, f(x_n) \approx y_n,$$

approximately. The type of function (for example, polynomials, exponential functions, sine and cosine functions) may be suggested by the nature of the problem (the underlying physical law, for instance), and in many cases a polynomial of a certain degree will be appropriate.

Let us begin with a motivation.

If we require strict equality $f(x_1) = y_1, \cdots, f(x_n) = y_n$ and use polynomials of sufficiently high degree, we may apply one of the methods discussed in Sec. 19.3 in connection with interpolation. However, in certain situations this would not be the appropriate solution of the actual problem. For instance, to the four points

(1) $$(-1.3, 0.103), \qquad (-0.1, 1.099), \qquad (0.2, 0.808), \qquad (1.3, 1.897)$$

there corresponds the interpolation polynomial $f(x) = x^3 - x + 1$ (Fig. 446), but if we graph the points, we see that they lie nearly on a straight line. Hence if these values are obtained in an experiment and thus involve an experimental error, and if the nature of the experiment suggests a linear relation, we better fit a straight line through the points (Fig. 446). Such a line may be useful for predicting values to be expected for other values of x. A widely used principle for fitting straight lines is the **method**

Fig. 446. Approximate fitting of a straight line

www.EngineeringEBooksPdf.com

of least squares by Gauss and Legendre. In the present situation it may be formulated as follows.

Method of Least Squares. *The straight line*

$$y = a + bx \tag{2}$$

should be fitted through the given points $(x_1, y_1), \cdots, (x_n, y_n)$ *so that the sum of the squares of the distances of those points from the straight line is minimum, where the distance is measured in the vertical direction (the y-direction).*

The point on the line with abscissa x_j has the ordinate $a + bx_j$. Hence its distance from (x_j, y_j) is $|y_j - a - bx_j|$ (Fig. 447) and that sum of squares is

$$q = \sum_{j=1}^{n} (y_j - a - bx_j)^2.$$

q depends on a and b. A necessary condition for q to be minimum is

$$\begin{aligned} \frac{\partial q}{\partial a} &= -2\sum (y_j - a - bx_j) = 0 \\ \frac{\partial q}{\partial b} &= -2\sum x_j (y_j - a - bx_j) = 0 \end{aligned} \tag{3}$$

(where we sum over j from 1 to n). Dividing by 2, writing each sum as three sums, and taking one of them to the right, we obtain the result

$$\begin{aligned} an \quad\quad + b\sum x_j &= \sum y_j \\ a\sum x_j + b\sum x_j^2 &= \sum x_j y_j. \end{aligned} \tag{4}$$

These equations are called the **normal equations** of our problem.

Fig. 447. Vetrical distance of a point (x_j, y_j) from a straight line $y = a + bx$

EXAMPLE 1 Straight Line

Using the method of least squares, fit a straight line to the four points given in formula (1).

Solution. We obtain

$$n = 4, \quad \sum x_j = 0.1, \quad \sum x_j^2 = 3.43, \quad \sum y_j = 3.907, \quad \sum x_j y_j = 2.3839.$$

www.EngineeringEBooksPdf.com

Hence the normal equations are

$$4a + 0.10b = 3.9070$$
$$0.1a + 3.43b = 2.3839.$$

The solution (rounded to 4D) is $a = 0.9601$, $b = 0.6670$, and we obtain the straight line (Fig. 446)

$$y = 0.9601 + 0.6670x.$$

Curve Fitting by Polynomials of Degree *m*

Our method of curve fitting can be generalized from a polynomial $y = a + bx$ to a polynomial of degree m

$$p(x) = b_0 + b_1x + \cdots + b_mx^m \tag{5}$$

where $m \leqq n - 1$. Then q takes the form

$$q = \sum_{j=1}^{n} (y_j - p(x_j))^2$$

and depends on $m + 1$ parameters $b_0, \cdots, b_m$. Instead of (3) we then have $m + 1$ conditions

$$\frac{\partial q}{\partial b_0} = 0, \qquad \cdots, \qquad \frac{\partial q}{\partial b_m} = 0 \tag{6}$$

which give a system of $m + 1$ normal equations.

In the case of a quadratic polynomial

$$p(x) = b_0 + b_1x + b_2x^2 \tag{7}$$

the normal equations are (summation from 1 to n)

$$\begin{aligned} b_0n \quad\quad + b_1\sum x_j + b_2\sum x_j^2 &= \sum y_j \\ b_0\sum x_j + b_1\sum x_j^2 + b_2\sum x_j^3 &= \sum x_jy_j \\ b_0\sum x_j^2 + b_1\sum x_j^3 + b_2\sum x_j^4 &= \sum x_j^2y_j. \end{aligned} \tag{8}$$

The derivation of (8) is left to the reader.

EXAMPLE 2 Quadratic Parabola by Least Squares

Fit a parabola through the data (0, 5), (2, 4), (4, 1), (6, 6), (8, 7).

Solution. For the normal equations we need $n = 5$, $\Sigma x_j = 20$, $\Sigma x_j^2 = 120$, $\Sigma x_j^3 = 800$, $\Sigma x_j^4 = 5664$, $\Sigma y_j = 23$, $\Sigma x_jy_j = 104$, $\Sigma x_j^2y_j = 696$. Hence these equations are

$$\begin{aligned} 5b_0 + 20b_1 + 120b_2 &= 23 \\ 20b_0 + 120b_1 + 800b_2 &= 104 \\ 120b_0 + 800b_1 + 5664b_2 &= 696. \end{aligned}$$

www.EngineeringEBooksPdf.com

Solving them we obtain the quadratic least squares parabola (Fig. 448)

$$y = 5.11429 - 1.41429x + 0.21429x^2.$$

Fig. 448. Least squares parabola in Example 2

For a general polynomial (5) the normal equations form a linear system of equations in the unknowns $b_0, \cdots, b_m$. When its matrix $\mathbf{M}$ is nonsingular, we can solve the system by Cholesky's method (Sec. 20.2) because then $\mathbf{M}$ is positive definite (and symmetric). When the equations are nearly linearly dependent, the normal equations may become ill-conditioned and should be replaced by other methods; see [E5], Sec. 5.7, listed in App. 1.

The least squares method also plays a role in statistics (see Sec. 25.9).

PROBLEM SET 20.5

1–6 FITTING A STRAIGHT LINE

Fit a straight line to the given points (x, y) by least squares. Show the details. Check your result by sketching the points and the line. Judge the goodness of fit.

1. (0, 2), (2, 0), (3, −2), (5, −3)

2. How does the line in Prob. 1 change if you add a point far above it, say, (1, 3)? Guess first.

3. (0, 1.8), (1, 1.6), (2, 1.1), (3, 1.5), (4, 2.3)

4. Hooke's law $F = ks$. Estimate the spring modulus k from the force F [lb] and the elongation s [cm], where $(F, s) = (1, 0.3)$, $(2, 0.7)$, $(4, 1.3)$, $(6, 1.9)$, $(10, 3.2)$, $(20, 6.3)$.

5. Average speed. Estimate the average speed v_{av} of a car traveling according to $s = v \cdot t$ [km] (s = distance traveled, t [hr] = time) from $(t, s) = (9, 140)$, $(10, 220)$, $(11, 310)$, $(12, 410)$.

6. Ohm's law $U = Ri$. Estimate R from $(i, U) = (2, 104)$, $(4, 206)$, $(6, 314)$, $(10, 530)$.

7. Derive the normal equations (8).

8–11 FITTING A QUADRATIC PARABOLA

Fit a parabola (7) to the points (x, y). Check by sketching.

8. (−1, 5), (1, 3), (2, 4), (3, 8)

9. (2, −3), (3, 0), (5, 1), (6, 0) (7, −2)

10. t [hr] = Worker's time on duty, y [sec] = His/her reaction time, $(t, y) = (1, 2.0)$, $(2, 1.78)$, $(3, 1.90)$, $(4, 2.35)$, $(5, 2.70)$

11. The data in Prob. 3. Plot the points, the line, and the parabola jointly. Compare and comment.

12. Cubic parabola. Derive the formula for the normal equations of a cubic least squares parabola.

13. Fit curves (2) and (7) and a cubic parabola by least squares to $(x, y) = (-2, -30)$, $(-1, -4)$, $(0, 4)$, $(1, 4)$, $(2, 22)$, $(3, 68)$. Graph these curves and the points on common axes. Comment on the goodness of fit.

14. TEAM PROJECT. The **least squares approximation of a function** $f(x)$ on an interval $a \leqq x \leqq b$ by a function

$$F_m(x) = a_0 y_0(x) + a_1 y_1(x) + \cdots + a_m y_m(x)$$

www.EngineeringEBooksPdf.com

where $y_0(x), \cdots, y_m(x)$ are given functions, requires the determination of the coefficients $a_0, \cdots, a_m$ such that

$$\int_a^b [f(x) - F_m(x)]^2 \, dx \tag{9}$$

becomes minimum. This integral is denoted by $\|f - F_m\|^2$, and $\|f - F_m\|$ is called the $\boldsymbol{L_2}$**-norm** of $f - F_m$ (L suggesting Lebesgue[5]). A necessary condition for that minimum is given by $\partial\|f - F_m\|^2/\partial a_j = 0$, $j = 0, \cdots, m$ [the analog of (6)]. **(a)** Show that this leads to $m + 1$ normal equations ($j = 0, \cdots, m$)

$$\sum_{k=0}^{m} h_{jk} a_k = b_j \qquad \text{where}$$

$$h_{jk} = \int_a^b y_j(x) y_k(x) \, dx, \tag{10}$$

$$b_j = \int_a^b f(x) y_j(x) \, dx.$$

(b) Polynomial. What form does (10) take if $F_m(x) = a_0 + a_1 x + \cdots + a_m x^m$? What is the coefficient matrix of (10) in this case when the interval is $0 \leqq x \leqq 1$?

(c) Orthogonal functions. What are the solutions of (10) if $y_0(x), \cdots, y_m(x)$ are orthogonal on the interval $a \leqq x \leqq b$? (For the definition, see Sec. 11.5. See also Sec. 11.6.)

15. CAS EXPERIMENT. Least Squares versus Interpolation. For the given data and for data of your choice find the interpolation polynomial and the least squares approximations (linear, quadratic, etc.). Compare and comment.

(a) $(-2, 0)$, $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(2, 0)$

(b) $(-4, 0)$, $(-3, 0)$, $(-2, 0)$, $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(2, 0)$, $(3, 0)$, $(4, 0)$

(c) Choose five points on a straight line, e.g., $(0, 0)$, $(1, 1), \cdots, (4, 4)$. Move one point 1 unit upward and find the quadratic least squares polynomial. Do this for each point. Graph the five polynomials on common axes. Which of the five motions has the greatest effect?

20.6 Matrix Eigenvalue Problems: Introduction

We now come to the second part of our chapter on numeric linear algebra. In the *first part of this chapter* we discussed methods of solving systems of linear equations, which included Gauss elimination with backward substitution. This method is known as a direct method since it gives solutions after a prescribed amount of computation. The Gauss method was modified by Doolittle's method, Crout's method, and Cholesky's method, each requiring fewer arithmetic operations than Gauss. Finally we presented indirect methods of solving systems of linear equations, that is, the Gauss–Seidel method and the Jacobi iteration. The indirect methods require an undetermined number of iterations. That number depends on how far we start from the true solution and what degree of accuracy we require. Moreover, depending on the problem, convergence may be fast or slow or our computation cycle might not even converge. This led to the concepts of ill-conditioned problems and condition numbers that help us gain some control over difficulties inherent in numerics.

The second part of this chapter deals with some of the most important ideas and numeric methods for matrix eigenvalue problems. This very extensive part of numeric linear algebra is of great practical importance, with much research going on, and hundreds, if not thousands, of papers published in various mathematical journals (see the references in [E8], [E9], [E11], [E29]). We begin with the concepts and general results we shall need in explaining and applying numeric methods for eigenvalue problems. (For typical models of eigenvalue problems see Chap. 8.)

[5]HENRI LEBESGUE (1875–1941), great French mathematician, creator of a modern theory of measure and integration in his famous doctoral thesis of 1902.

www.EngineeringEBooksPdf.com

An **eigenvalue** or **characteristic value** (or *latent root*) of a given $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is a real or complex number λ such that the vector equation

$$\mathbf{Ax} = \lambda\mathbf{x} \tag{1}$$

has a nontrivial solution, that is, a solution $\mathbf{x} \neq \mathbf{0}$, which is then called an **eigenvector** or **characteristic vector** of $\mathbf{A}$ corresponding to that eigenvalue λ. The set of all eigenvalues of $\mathbf{A}$ is called the **spectrum** of $\mathbf{A}$. Equation (1) can be written

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0} \tag{2}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix. This homogeneous system has a nontrivial solution if and only if the **characteristic determinant** det $(\mathbf{A} - \lambda\mathbf{I})$ is 0 (see Theorem 2 in Sec. 7.5). This gives (see Sec. 8.1)

THEOREM 1

Eigenvalues

The eigenvalues of $\mathbf{A}$ *are the solutions* λ *of the* **characteristic equation**

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{3}$$

Developing the characteristic determinant, we obtain the **characteristic polynomial** of $\mathbf{A}$, which is of degree n in λ. Hence $\mathbf{A}$ has at least one and at most n numerically different eigenvalues. If $\mathbf{A}$ is real, so are the coefficients of the characteristic polynomial. By familiar algebra it follows that then the roots (the eigenvalues of $\mathbf{A}$) are ***real or complex conjugates*** in pairs.

To give you some orientation of the underlying approaches of numerics for eigenvalue problems, note the following. For large or very large matrices it may be very difficult to determine the eigenvalues, since, in general, it is difficult to find the roots of characteristic polynomials of higher degrees. We will discuss different numeric methods for finding eigenvalues that achieve different results. Some methods, such as in Sec. 20.7, will give us only regions in which complex eigenvalues lie (Geschgorin's method) or the intervals in which the largest and smallest real eigenvalue lie (Collatz method). Other methods compute all eigenvalues, such as the Householder tridiagonalization method and the QR-method in Sec. 20.9.

To continue our discussion, we shall usually denote the eigenvalues of $\mathbf{A}$ by

$$\lambda_1, \lambda_2, \cdots, \lambda_n$$

with the understanding that some (or all) of them may be equal.

The sum of these n eigenvalues equals the sum of the entries on the main diagonal of $\mathbf{A}$, called the trace of $\mathbf{A}$; thus

$$\text{trace } \mathbf{A} = \sum_{j=1}^{n} a_{jj} = \sum_{k=1}^{n} \lambda_k. \tag{4}$$

www.EngineeringEBooksPdf.com

Also, the product of the eigenvalues equals the determinant of **A**,

$$\det \mathbf{A} = \lambda_1 \lambda_2 \cdots \lambda_n. \tag{5}$$

Both formulas follow from the product representation of the characteristic polynomial, which we denote by $f(\lambda)$,

$$f(\lambda) = (-1)^n (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n).$$

If we take equal factors together and denote the *numerically distinct* eigenvalues of **A** by $\lambda_1, \cdots, \lambda_r$ $(r \leqq n)$, then the product becomes

$$f(\lambda) = (-1)^n (\lambda - \lambda_1)^{m_1} (\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_r)^{m_r}. \tag{6}$$

The exponent m_j is called the **algebraic multiplicity** of λ_j. The maximum number of linearly independent eigenvectors corresponding to λ_j is called the **geometric multiplicity** of λ_j. It is equal to or smaller than m_j.

A subspace S of R^n or C^n (if **A** is complex) is called an **invariant subspace** of **A** if for every **v** in S the vector **Av** is also in S. **Eigenspaces** of **A** (spaces of eigenvectors; Sec. 8.1) are important invariant subspaces of **A**.

An $n \times n$ matrix **B** is called **similar** to **A** if there is a nonsingular $n \times n$ matrix **T** such that

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}. \tag{7}$$

Similarity is important for the following reason.

THEOREM 2 **Similar Matrices**

Similar matrices have the same eigenvalues. If **x** *is an eigenvector of* **A**, *then* $\mathbf{y} = \mathbf{T}^{-1}\mathbf{x}$ *is an eigenvector of* **B** *in* (7) *corresponding to the same eigenvalue.* (Proof in Sec. 8.4.)

Another theorem that has various applications in numerics is as follows.

THEOREM 3 **Spectral Shift**

If **A** *has the eigenvalues* $\lambda_1, \cdots, \lambda_n$, *then* $\mathbf{A} - k\mathbf{I}$ *with arbitrary* k *has the eigenvalues* $\lambda_1 - k, \cdots, \lambda_n - k$.

This theorem is a special case of the following **spectral mapping theorem**.

THEOREM 4 **Polynomial Matrices**

If λ *is an eigenvalue of* **A**, *then*

$$q(\lambda) = \alpha_s \lambda^s + \alpha_{s-1}\lambda^{s-1} + \cdots + \alpha_1 \lambda + \alpha_0$$

is an eigenvalue of the **polynomial matrix**

$$q(\mathbf{A}) = \alpha_s \mathbf{A}^s + \alpha_{s-1}\mathbf{A}^{s-1} + \cdots + \alpha_1 \mathbf{A} + a_0 \mathbf{I}.$$

www.EngineeringEBooksPdf.com

PROOF $\mathbf{Ax} = \lambda\mathbf{x}$ implies $\mathbf{A}^2\mathbf{x} = \mathbf{A}\lambda\mathbf{x} = \lambda\mathbf{Ax} = \lambda^2\mathbf{x}$, $\mathbf{A}^3\mathbf{x} = \lambda^3\mathbf{x}$, etc. Thus

$$\begin{aligned} q(\mathbf{A})\mathbf{x} &= (\alpha_s\mathbf{A}^s + \alpha_{s-1}\mathbf{A}^{s-1} + \cdots)\,\mathbf{x} \\ &= \alpha_s\mathbf{A}^s\mathbf{x} + \alpha_{s-1}A^{s-1}\mathbf{x} + \cdots \\ &= \alpha_s\lambda^s\mathbf{x} + \alpha_{s-1}\lambda^{s-1}\mathbf{x} + \cdots = q(\lambda)\,\mathbf{x}. \end{aligned}$$

■

The eigenvalues of important special matrices can be characterized as follows.

THEOREM 5 **Special Matrices**

The eigenvalues of Hermitian matrices (i.e., $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}$), *hence of real symmetric matrices* (i.e., $\mathbf{A}^{\mathsf{T}} = \mathbf{A}$), *are real. The eigenvalues of skew-Hermitian matrices* (i.e., $\overline{\mathbf{A}}^{\mathsf{T}} = -\mathbf{A}$), *hence of real skew-symmetric matrices* (i.e., $\mathbf{A}^{\mathsf{T}} = -\mathbf{A}$), *are pure imaginary or* 0. *The eigenvalues of unitary matrices* (i.e., $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$), *hence of orthogonal matrices* (i.e., $\mathbf{A}^{\mathsf{T}} = \mathbf{A}^{-1}$), *have absolute value* 1. (Proofs in Secs. 8.3 and 8.5.)

The **choice of a numeric method** for matrix eigenvalue problems depends essentially on two circumstances, on the kind of matrix (real symmetric, real general, complex, sparse, or full) and on the kind of information to be obtained, that is, whether one wants to know all eigenvalues or merely specific ones, for instance, the largest eigenvalue, whether eigenvalues *and* eigenvectors are wanted, and so on. It is clear that we cannot enter into a systematic discussion of all these and further possibilities that arise in practice, but we shall concentrate on some basic aspects and methods that will give us a general understanding of this fascinating field.

20.7 Inclusion of Matrix Eigenvalues

The whole of numerics for matrix eigenvalues is motivated by the fact that, except for a few trivial cases, we cannot determine eigenvalues *exactly* by a finite process because these values are the roots of a polynomial of nth degree. Hence we must mainly use iteration.

In this section we state a few general theorems that give approximations and error bounds for eigenvalues. Our matrices will continue to be real (except in formula (5) below), but since (nonsymmetric) matrices may have complex eigenvalues, complex numbers will play a (very modest) role in this section.

The important theorem by Gerschgorin gives a region consisting of closed circular disks in the complex plane and including all the eigenvalues of a given matrix. Indeed, for each $j = 1, \cdots, n$ the inequality (1) in the theorem determines a closed circular disk in the complex λ-plane with center a_{jj} and radius given by the right side of (1); and Theorem 1 states that each of the eigenvalues of $\mathbf{A}$ lies in one of these n disks.

THEOREM 1 **Gerschgorin's Theorem**[6]

Let λ be an eigenvalue of an arbitrary $n \times n$ matrix $\mathbf{A} = [a_{jk}]$. Then for some integer j $(1 \leqq j \leqq n)$ we have

$$(1) \qquad |a_{jj} - \lambda| \leqq |a_{j1}| + |a_{j2}| + \cdots + |a_{j,j-1}| + |a_{j,j+1}| + \cdots + |a_{jn}|.$$

[6]SEMYON ARANOVICH GERSCHGORIN (1901–1933), Russian mathematician.

www.EngineeringEBooksPdf.com

PROOF Let $\mathbf{x}$ be an eigenvector corresponding to an eigenvalue λ of $\mathbf{A}$. Then

$$\mathbf{Ax} = \lambda\mathbf{x} \qquad \text{or} \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \tag{2}$$

Let x_j be a component of $\mathbf{x}$ that is largest in absolute value. Then we have $|x_m/x_j| \leqq 1$ for $m = 1, \cdots, n$. The vector equation (2) is equivalent to a system of n equations for the n components of the vectors on both sides. The jth of these n equations with j as just indicated is

$$a_{j1}x_1 + \cdots + a_{j,j-1}x_{j-1} + (a_{jj} - \lambda)x_j + a_{j,j+1}x_{j+1} + \cdots + a_{jn}x_n = 0.$$

Division by x_j (which cannot be zero; why?) and reshuffling terms gives

$$a_{jj} - \lambda = -a_{j1}\frac{x_1}{x_j} - \cdots - a_{j,j-1}\frac{x_{j-1}}{x_j} - a_{j,j+1}\frac{x_{j+1}}{x_j} - \cdots - a_{jn}\frac{x_n}{x_j}.$$

By taking absolute values on both sides of this equation, applying the triangle inequality $|a + b| \leqq |a| + |b|$ (where a and b are any complex numbers), and observing that because of the choice of j (which is crucial!), $|x_1/x_j| \leqq 1, \cdots, |x_n/x_j| \leqq 1$, we obtain (1), and the theorem is proved. ■

EXAMPLE 1 Gerschgorin's Theorem

For the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 5 & 1 \\ \frac{1}{2} & 1 & 1 \end{bmatrix}$$

we get the Gerschgorin disks (Fig. 449)

D_1: Center 0, radius 1, D_2: Center 5, radius 1.5, D_3: Center 1, radius 1.5.

The centers are the main diagonal entries of $\mathbf{A}$. These would be the eigenvalues of $\mathbf{A}$ if $\mathbf{A}$ were diagonal. We can take these values as crude approximations of the unknown eigenvalues (3D-values) $\lambda_1 = -0.209$, $\lambda_2 = 5.305$, $\lambda_3 = 0.904$ (verify this); then the radii of the disks are corresponding error bounds.

Since $\mathbf{A}$ is symmetric, it follows from Theorem 5, Sec. 20.6, that the spectrum of $\mathbf{A}$ must actually lie in the intervals $[-1, 2.5]$ and $[3.5, 6.5]$.

It is interesting that here the Gerschgorin disks form two disjoint sets, namely, $D_1 \cup D_3$, which contains two eigenvalues, and D_2, which contains one eigenvalue. This is typical, as the following theorem shows. ■

Fig. 449. Gerschgorin disks in Example 1

www.EngineeringEBooksPdf.com

THEOREM 2 **Extension of Gerschgorin's Theorem**

If p Gerschgorin disks form a set S that is disjoint from the $n - p$ other disks of a given matrix $\mathbf{A}$, *then S contains precisely p eigenvalues of* $\mathbf{A}$ (*each counted with its algebraic multiplicity, as defined in Sec.* 20.6).

Idea of Proof. Set $\mathbf{A} = \mathbf{B} + \mathbf{C}$, where $\mathbf{B}$ is the diagonal matrix with entries a_{jj}, and apply Theorem 1 to $\mathbf{A}_t = \mathbf{B} + t\mathbf{C}$ with real t growing from 0 to 1. ■

EXAMPLE 2 **Another Application of Gerschgorin's Theorem. Similarity**

Suppose that we have diagonalized a matrix by some numeric method that left us with some off-diagonal entries of size 10^{-5}, say,

$$\mathbf{A} = \begin{bmatrix} 2 & 10^{-5} & 10^{-5} \\ 10^{-5} & 2 & 10^{-5} \\ 10^{-5} & 10^{-5} & 4 \end{bmatrix}.$$

What can we conclude about deviations of the eigenvalues from the main diagonal entries?

Solution. By Theorem 2, one eigenvalue must lie in the disk of radius $2 \cdot 10^{-5}$ centered at 4 and two eigenvalues (or an eigenvalue of algebraic multiplicity 2) in the disk of radius $2 \cdot 10^{-5}$ centered at 2. Actually, since the matrix is symmetric, these eigenvalues must lie in the intersections of these disks and the real axis, by Theorem 5 in Sec. 20.6.

We show how an isolated disk can always be reduced in size by a similarity transformation. The matrix

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{-5} \end{bmatrix} \begin{bmatrix} 2 & 10^{-5} & 10^{-5} \\ 10^{-5} & 2 & 10^{-5} \\ 10^{-5} & 10^{-5} & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{5} \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 10^{-5} & 1 \\ 10^{-5} & 2 & 1 \\ 10^{-10} & 10^{-10} & 4 \end{bmatrix}$$

is similar to $\mathbf{A}$. Hence by Theorem 2, Sec. 20.6, it has the same eigenvalues as $\mathbf{A}$. From Row 3 we get the smaller disk of radius $2 \cdot 10^{-10}$. Note that the other disks got bigger, approximately by a factor of 10^5. And in choosing $\mathbf{T}$ we have to watch that the new disks do not overlap with the disk whose size we want to decrease.

For further interesting facts, see the book [E28]. ■

By definition, a **diagonally dominant** matrix $\mathbf{A} = [a_{jk}]$ is an $n \times n$ matrix such that

$$|a_{jj}| \geqq \sum_{k \neq j} |a_{jk}| \qquad j = 1, \cdots, n \tag{3}$$

where we sum over all off-diagonal entries in Row j. The matrix is said to be **strictly diagonally dominant** if $>$ in (3) for all j. Use Theorem 1 to prove the following basic property.

THEOREM 3 **Strict Diagonal Dominance**

Strictly diagonally dominant matrices are nonsingular.

www.EngineeringEBooksPdf.com

Further Inclusion Theorems

An **inclusion theorem** is a theorem that specifies a set which contains at least one eigenvalue of a given matrix. Thus, Theorems 1 and 2 are inclusion theorems; they even include the whole spectrum. We now discuss some famous theorems that yield further inclusions of eigenvalues. We state the first two of them without proofs (which would exceed the level of this book).

THEOREM 4

Schur's Theorem[7]

Let $\mathbf{A} = [a_{jk}]$ *be a* $n \times n$ *matrix. Then for each of its eigenvalues* $\lambda_1, \cdots, \lambda_n$,

$$|\lambda_m|^2 \leqq \sum_{i=1}^{n} |\lambda_i|^2 \leqq \sum_{j=1}^{n} \sum_{k=1}^{n} |a_{jk}|^2 \quad \textbf{(Schur's inequality)}. \tag{4}$$

In (4) *the second equality sign holds if and only if* $\mathbf{A}$ *is such that*

$$\overline{\mathbf{A}}^{\mathsf{T}}\mathbf{A} = \mathbf{A}\overline{\mathbf{A}}^{\mathsf{T}}. \tag{5}$$

Matrices that satisfy (5) are called **normal matrices**. It is not difficult to see that Hermitian, skew-Hermitian, and unitary matrices are normal, and so are real symmetric, skew-symmetric, and orthogonal matrices.

EXAMPLE 3 **Bounds for Eigenvalues Obtained from Schur's Inequality**

For the matrix

$$\mathbf{A} = \begin{bmatrix} 26 & -2 & 2 \\ 2 & 21 & 4 \\ 4 & 2 & 28 \end{bmatrix}$$

we obtain from Schur's inequality $|\lambda| \leqq \sqrt{1949} = 44.1475$. You may verify that the eigenvalues are 30, 25, and 20. Thus $30^2 + 25^2 + 20^2 = 1925 < 1949$; in fact, $\mathbf{A}$ is not normal. ■

The preceding theorems are valid for every real or complex square matrix. Other theorems hold for special classes of matrices only. Famous is the following one, which has various applications, for instance, in economics.

THEOREM 5

Perron's Theorem[8]

Let $\mathbf{A}$ *be a real* $n \times n$ *matrix whose entries are all positive. Then* $\mathbf{A}$ *has a positive real eigenvalue* $\lambda = \rho$ *of multiplicity* 1. *The corresponding eigenvector can be chosen with all components positive.* (*The other eigenvalues are less than* ρ *in absolute value.*)

[7]ISSAI SCHUR (1875–1941), German mathematician, also known by his important work in group theory.

[8]OSKAR PERRON (1880–1975) and GEORG FROBENIUS (1849–1917), German mathematicians, known for their work in potential theory, ODEs (Sec. 5.4), and group theory.

www.EngineeringEBooksPdf.com

For a proof see Ref. [B3], vol. II, pp. 53–62. The theorem also holds for matrices with *nonnegative* real entries ("**Perron–Frobenius Theorem**"[8]) provided **A** is **irreducible**, that is, it cannot be brought to the following form by interchanging rows and columns; here **B** and **F** are square and **0** is a zero matrix.

$$\begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{0} & \mathbf{F} \end{bmatrix}$$

Perron's theorem has various applications, for instance, in economics. It is interesting that one can obtain from it a theorem that gives a numeric algorithm:

THEOREM 6

Collatz Inclusion Theorem[9]

Let $\mathbf{A} = [a_{jk}]$ *be a real* $n \times n$ *matrix whose elements are all positive. Let* $\mathbf{x}$ *be any real vector whose components* $x_1, \cdots, x_n$ *are positive, and let* $y_1, \cdots, y_n$ *be the components of the vector* $\mathbf{y} = \mathbf{Ax}$. *Then the closed interval on the real axis bounded by the smallest and the largest of the* n *quotients* $q_j = y_j/x_j$ *contains at least one eigenvalue of* $\mathbf{A}$.

PROOF We have $\mathbf{Ax} = \mathbf{y}$ or

$$\mathbf{y} - \mathbf{Ax} = \mathbf{0}. \tag{6}$$

The transpose $\mathbf{A}^\mathsf{T}$ satisfies the conditions of Theorem 5. Hence $\mathbf{A}^\mathsf{T}$ has a positive eigenvalue λ and, corresponding to this eigenvalue, an eigenvector $\mathbf{u}$ whose components u_j are all positive. Thus $\mathbf{A}^\mathsf{T}\mathbf{u} = \lambda\mathbf{u}$ and by taking the transpose we obtain $\mathbf{u}^\mathsf{T}\mathbf{A} = \lambda\mathbf{u}^\mathsf{T}$. From this and (6) we have

$$\mathbf{u}^\mathsf{T}(\mathbf{y} - \mathbf{Ax}) = \mathbf{u}^\mathsf{T}\mathbf{y} - \mathbf{u}^\mathsf{T}\mathbf{Ax} = \mathbf{u}^\mathsf{T}\mathbf{y} - \lambda\mathbf{u}^\mathsf{T}\mathbf{x} = \mathbf{u}^\mathsf{T}(\mathbf{y} - \lambda\mathbf{x}) = 0$$

or written out

$$\sum_{j=1}^{n} u_j(y_j - \lambda x_j) = 0.$$

Since all the components u_j are positive, it follows that

$$\begin{aligned} &y_j - \lambda x_j \geqq 0, \quad \text{that is,} \quad q_j \geqq \lambda \quad \text{for at least one } j, \\ &y_j - \lambda x_j \leqq 0, \quad \text{that is,} \quad q_j \leqq \lambda \quad \text{for at least one } j. \end{aligned} \quad \text{and} \tag{7}$$

Since $\mathbf{A}$ and $\mathbf{A}^\mathsf{T}$ have the same eigenvalues, λ is an eigenvalue of $\mathbf{A}$, and from (7) the statement of the theorem follows. ■

[9]LOTHAR COLLATZ (1910–1990), German mathematician known for his work in numerics.

www.EngineeringEBooksPdf.com

EXAMPLE 4 Bounds for Eigenvalues from Collatz's Theorem. Iteration

For a given matrix $\mathbf{A}$ with positive entries we choose an $\mathbf{x} = \mathbf{x}_0$ and **iterate**, that is, we compute $\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0$, $\mathbf{x}_2 = \mathbf{A}\mathbf{x}_1, \cdots, \mathbf{x}_{20} = \mathbf{A}\mathbf{x}_{19}$. In each step, taking $\mathbf{x} = \mathbf{x}_j$ and $\mathbf{y} = \mathbf{A}\mathbf{x}_j = \mathbf{x}_{j+1}$ we compute an inclusion interval by Collatz's theorem. This gives (6S)

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.02 & 0.22 \\ 0.02 & 0.28 & 0.20 \\ 0.22 & 0.20 & 0.40 \end{bmatrix}, \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{x}_1 = \begin{bmatrix} 0.73 \\ 0.50 \\ 0.82 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0.5481 \\ 0.3186 \\ 0.5886 \end{bmatrix},$$

$$\cdots, \mathbf{x}_{19} = \begin{bmatrix} 0.00216309 \\ 0.00108155 \\ 0.00216309 \end{bmatrix}, \mathbf{x}_{20} = \begin{bmatrix} 0.00155743 \\ 0.000778713 \\ 0.00155743 \end{bmatrix}$$

and the intervals $0.5 \leqq \lambda \leqq 0.82$, $0.3186/0.50 = 0.6372 \leqq \lambda \leqq 0.5481/0.73 = 0.750822$, etc. These intervals have length

j	1	2	3	10	15	20
Length	0.32	0.113622	0.0539835	0.0004217	0.0000132	0.0000004

Using the characteristic polynomial, you may verify that the eigenvalues of $\mathbf{A}$ are 0.72, 0.36, 0.09, so that those intervals include the largest eigenvalue, 0.72. Their lengths decreased with j, so that the iteration was worthwhile. The reason will appear in the next section, where we discuss an iteration method for eigenvalues. ■

PROBLEM SET 20.7

1–6 GERSCHGORIN DISKS

Find and sketch disks or intervals that contain the eigenvalues. If you have a CAS, find the spectrum and compare.

1. $\begin{bmatrix} 5 & 2 & 4 \\ -2 & 0 & 2 \\ 2 & 4 & 7 \end{bmatrix}$ **2.** $\begin{bmatrix} 5 & 10^{-2} & 10^{-2} \\ 10^{-2} & 8 & 10^{-2} \\ 10^{-2} & 10^{-2} & 9 \end{bmatrix}$

3. $\begin{bmatrix} 0 & 0.4 & -0.1 \\ -0.4 & 0 & 0.3 \\ 0.1 & -0.3 & 0 \end{bmatrix}$ **4.** $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 4 & 3 \\ 1 & 3 & 12 \end{bmatrix}$

5. $\begin{bmatrix} 2 & i & 1+i \\ -i & 3 & 0 \\ 1-i & 0 & 8 \end{bmatrix}$ **6.** $\begin{bmatrix} 10 & 0.1 & -0.2 \\ 0.1 & 6 & 0 \\ -0.2 & 0 & 3 \end{bmatrix}$

7. Similarity. In Prob. 2, find $\mathbf{T}^{-\mathsf{T}}\mathbf{A}\mathbf{T}$ such that the radius of the Gerschgorin circle with center 5 is reduced by a factor 1/100.

8. By what integer factor can you at most reduce the Gerschgorin circle with center 3 in Prob. 6?

9. If a symmetric $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ has been diagonalized except for small off-diagonal entries of size 10^{-5}, what can you say about the eigenvalues?

10. Optimality of Gerschgorin disks. Illustrate with a 2×2 matrix that an eigenvalue may very well lie on a Gerschgorin circle, so that Gerschgorin disks can generally not be replaced with smaller disks without losing the inclusion property.

11. Spectral radius $\rho(\mathbf{A})$. Using Theorem 1, show that $\rho(\mathbf{A})$ cannot be greater than the row sum norm of $\mathbf{A}$.

12–16 SPECTRAL RADIUS

Use (4) to obtain an upper bound for the spectral radius:

12. In Prob. 4 **13.** In Prob. 1

14. In Prob. 6 **15.** In Prob. 3

16. In Prob. 5

17. Verify that the matrix in Prob. 5 is normal.

18. Normal matrices. Show that Hermitian, skew-Hermitian, and unitary matrices (hence real symmetric, skew-symmetric, and orthogonal matrices) are normal. Why is this of practical interest?

19. Prove Theorem 3 by using Theorem 1.

20. Extended Gerschgorin theorem. Prove Theorem 2. *Hint.* Let $\mathbf{A} = \mathbf{B} + \mathbf{C}$, $\mathbf{B} = \text{diag}\,(a_{jj})$, $\mathbf{A}_t = \mathbf{B} + t\mathbf{C}$, and let t increase continuously from 0 to 1.

www.EngineeringEBooksPdf.com

20.8 Power Method for Eigenvalues

A simple standard procedure for computing approximate values of the eigenvalues of an $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the **power method**. In this method we start from any vector $\mathbf{x}_0\ (\neq \mathbf{0})$ with n components and compute successively

$$\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0, \qquad \mathbf{x}_2 = \mathbf{A}\mathbf{x}_1, \qquad \cdots, \qquad \mathbf{x}_s = \mathbf{A}\mathbf{x}_{s-1}.$$

For simplifying notation, we denote $\mathbf{x}_{s-1}$ by $\mathbf{x}$ and $\mathbf{x}_s$ by $\mathbf{y}$, so that $\mathbf{y} = \mathbf{A}\mathbf{x}$.

The method applies to any $n \times n$ matrix $\mathbf{A}$ that has a **dominant eigenvalue** (a λ such that $|\lambda|$ is greater than the absolute values of the other eigenvalues). If $\mathbf{A}$ is *symmetric*, it also gives the error bound (2), in addition to the approximation (1).

THEOREM 1

Power Method, Error Bounds

Let $\mathbf{A}$ be an $n \times n$ real symmetric matrix. Let $\mathbf{x}\ (\neq \mathbf{0})$ be any real vector with n components. Furthermore, let

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \qquad m_0 = \mathbf{x}^\mathsf{T}\mathbf{x}, \qquad m_1 = \mathbf{x}^\mathsf{T}\mathbf{y}, \qquad m_2 = \mathbf{y}^\mathsf{T}\mathbf{y}.$$

Then the quotient

$$q = \frac{m_1}{m_0} \tag{1}$$

(Rayleigh[10] quotient)

is an approximation for an eigenvalue λ of $\mathbf{A}$ (usually that which is greatest in absolute value, but no general statements are possible).

Furthermore, if we set $q = \lambda - \epsilon$, so that ϵ is the error of q, then

$$|\epsilon| \leqq \delta = \sqrt{\frac{m_2}{m_0} - q^2}. \tag{2}$$

PROOF δ^2 denotes the radicand in (2). Since $m_1 = qm_0$ by (1), we have

$$(\mathbf{y} - q\mathbf{x})^\mathsf{T}(\mathbf{y} - q\mathbf{x}) = m_2 - 2qm_1 + q^2 m_0 = m_2 - q^2 m_0 = \delta^2 m_0. \tag{3}$$

Since $\mathbf{A}$ is real symmetric, it has an orthogonal set of n real unit eigenvectors $\mathbf{z}_1, \cdots, \mathbf{z}_n$ corresponding to the eigenvalues $\lambda_1, \cdots, \lambda_n$, respectively (some of which may be equal). (Proof in Ref. [B3], vol. 1, pp. 270–272, listed in App. 1.) Then $\mathbf{x}$ has a representation of the form

$$\mathbf{x} = a_1\mathbf{z}_1 + \cdots + a_n\mathbf{z}_n.$$

[10] LORD RAYLEIGH (JOHN WILLIAM STRUTT) (1842–1919), great English physicist and mathematician, professor at Cambridge and London, known for his important contributions to various branches of applied mathematics and theoretical physics, in particular, the theory of waves, elasticity, and hydrodynamics. In 1904 he received a Nobel Prize in physics.

www.EngineeringEBooksPdf.com

Now $\mathbf{Az}_1 = \lambda_1\mathbf{z}_1$, etc., and we obtain

$$\mathbf{y} = \mathbf{Ax} = a_1\lambda_1\mathbf{z}_1 + \cdots + a_n\lambda_n\mathbf{z}_n$$

and, since the $\mathbf{z}_j$ are orthogonal unit vectors,

$$m_0 = \mathbf{x}^\mathsf{T}\mathbf{x} = a_1^2 + \cdots + a_n^2. \tag{4}$$

It follows that in (3),

$$\mathbf{y} - q\mathbf{x} = a_1(\lambda_1 - q)\mathbf{z}_1 + \cdots + a_n(\lambda_n - q)\mathbf{z}_n.$$

Since the $\mathbf{z}_j$ are orthogonal unit vectors, we thus obtain from (3)

$$\delta^2 m_0 = (y - q\mathbf{x})^\mathsf{T}(\mathbf{y} - q\mathbf{x}) = a_1^2(\lambda_1 - q)^2 + \cdots + a_n^2(\lambda_n - q)^2. \tag{5}$$

Now let λ_c be an eigenvalue of $\mathbf{A}$ to which q is closest, where c suggests "closest." Then $(\lambda_c - q)^2 \leqq (\lambda_j - q)^2$ for $j = 1, \cdots, n$. From this and (5) we obtain the inequality

$$\delta^2 m_0 \geqq (\lambda_c - q)^2(a_1^2 + \cdots + a_n^2) = (\lambda_c - q)^2 m_0.$$

Dividing by m_0, taking square roots, and recalling the meaning of δ^2 gives

$$\delta = \sqrt{\frac{m_2}{m_0} - q^2} \geqq |\lambda_c - q|.$$

This shows that δ is a bound for the error ϵ of the approximation q of an eigenvalue of $\mathbf{A}$ and completes the proof. ■

The main advantage of the method is its simplicity. And it can handle *sparse matrices* too large to store as a full square array. Its disadvantage is its possibly slow convergence. From the proof of Theorem 1 we see that the speed of convergence depends on the ratio of the dominant eigenvalue to the next in absolute value (2:1 in Example 1, below).

If we want a convergent sequence of **eigenvectors,** then at the beginning of each step we **scale** the vector, say, by dividing its components by an absolutely largest one, as in Example 1, as follows.

EXAMPLE 1 Application of Theorem 1. Scaling

For the symmetric matrix $\mathbf{A}$ in Example 4, Sec. 20.7, and $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$ we obtain from (1) and (2) and the indicated scaling

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.02 & 0.22 \\ 0.02 & 0.28 & 0.20 \\ 0.22 & 0.20 & 0.40 \end{bmatrix}, \quad \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 0.890244 \\ 0.609756 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0.931193 \\ 0.541284 \\ 1 \end{bmatrix}$$

$$\mathbf{x}_5 = \begin{bmatrix} 0.990663 \\ 0.504682 \\ 1 \end{bmatrix}, \quad \mathbf{x}_{10} = \begin{bmatrix} 0.999707 \\ 0.500146 \\ 1 \end{bmatrix}, \quad \mathbf{x}_{15} = \begin{bmatrix} 0.999991 \\ 0.500005 \\ 1 \end{bmatrix}.$$

www.EngineeringEBooksPdf.com

Here $\mathbf{Ax_0} = [0.73 \quad 0.5 \quad 0.82]^\mathsf{T}$, scaled to $\mathbf{x_1} = [0.73/0.82 \quad 0.5/0.82 \quad 1]^\mathsf{T}$, etc. The dominant eigenvalue is 0.72, an eigenvector $[1 \quad 0.5 \quad 1]^\mathsf{T}$. The corresponding q and δ are computed each time before the next scaling. Thus in the first step,

$$q = \frac{m_1}{m_0} = \frac{\mathbf{x}_0^\mathsf{T}\mathbf{Ax}_0}{\mathbf{x}_0^\mathsf{T}\mathbf{x}_0} = \frac{2.05}{3} = 0.683333$$

$$\delta = \left(\frac{m_2}{m_0} - q^2\right)^{1/2} = \left(\frac{(\mathbf{Ax}_0)^\mathsf{T}\mathbf{Ax}_0}{\mathbf{x}_0^\mathsf{T}\mathbf{x}_0} - q^2\right)^{1/2} = \left(\frac{1.4553}{3} - q^2\right)^{1/2} = 0.134743.$$

This gives the following values of q, δ, and the error $\epsilon = 0.72 - q$ (calculations with 10D, rounded to 6D):

j	1	2	5	10
q	0.683333	0.716048	0.719944	0.720000
δ	0.134743	0.038887	0.004499	0.000141
ϵ	0.036667	0.003952	0.000056	$5 \cdot 10^{-8}$

The error bounds are much larger than the actual errors. This is typical, although the bounds cannot be improved; that is, for special symmetric matrices they agree with the errors.

Our present results are somewhat better than those of Collatz's method in Example 4 of Sec. 20.7, at the expense of more operations. ■

Spectral shift, the transition from $\mathbf{A}$ to $\mathbf{A} - k\mathbf{I}$, shifts every eigenvalue by $-k$. Although finding a good k can hardly be made automatic, it may be helped by some other method or small preliminary computational experiments. In Example 1, Gerschgorin's theorem gives $-0.02 \leqq \lambda \leqq 0.82$ for the whole spectrum (verify!). Shifting by -0.4 might be too much (then $-0.42 \leqq \lambda \leqq 0.42$), so let us try -0.2.

EXAMPLE 2 Power Method with Spectral Shift

For $\mathbf{A} - 0.2\mathbf{I}$ with $\mathbf{A}$ as in Example 1 we obtain the following substantial improvements (where the index 1 refers to Example 1 and the index 2 to the present example).

j	1	2	5	10
δ_1	0.134743	0.038887	0.004499	0.000141
δ_2	0.134743	0.034474	0.000693	$1.8 \cdot 10^{-6}$
ϵ_1	0.036667	0.003952	0.000056	$5 \cdot 10^{-8}$
ϵ_2	0.036667	0.002477	$1.3 \cdot 10^{-6}$	$9 \cdot 10^{-12}$

■

PROBLEM SET 20.8

1–4 POWER METHOD WITHOUT SCALING

Apply the power method without scaling (3 steps), using $\mathbf{x}_0 = [1, \quad 1]^\mathsf{T}$ or $[1 \quad 1 \quad 1]^\mathsf{T}$. Give Rayleigh quotients and error bounds. Show the details of your work.

1. $\begin{bmatrix} 9 & 4 \\ 4 & 3 \end{bmatrix}$

2. $\begin{bmatrix} 7 & -3 \\ -3 & -1 \end{bmatrix}$

3. $\begin{bmatrix} 2 & -1 & 1 \\ -1 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix}$

4. $\begin{bmatrix} 3.6 & -1.8 & 1.8 \\ -1.8 & 2.8 & -2.6 \\ 1.8 & -2.6 & 2.8 \end{bmatrix}$

5–8 POWER METHOD WITH SCALING

Apply the power method (3 steps) with scaling, using $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$ or $[1 \quad 1 \quad 1 \quad 1]^\mathsf{T}$, as applicable. Give

www.EngineeringEBooksPdf.com

Rayleigh quotients and error bounds. Show the details of your work.

5. The matrix in Prob. 3

6. $\begin{bmatrix} 4 & 2 & 3 \\ 2 & 7 & 6 \\ 3 & 6 & 4 \end{bmatrix}$

7. $\begin{bmatrix} 5 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & 5 \end{bmatrix}$

8. $\begin{bmatrix} 2 & 4 & 0 & 1 \\ 4 & 1 & 2 & 8 \\ 0 & 2 & 5 & 2 \\ 1 & 8 & 2 & 0 \end{bmatrix}$

9. Prove that if $\mathbf{x}$ is an eigenvector, then $\delta = 0$ in (2). Give two examples.

10. Rayleigh quotient. Why does q generally approximate the eigenvalue of greatest absolute value? When will q be a good approximation?

11. Spectral shift, smallest eigenvalue. In Prob. 3 set $\mathbf{B} = \mathbf{A} - 3\mathbf{I}$ (as perhaps suggested by the diagonal entries) and see whether you may get a sequence of q's converging to an eigenvalue of $\mathbf{A}$ that *is smallest* (not largest) in absolute value. Use $\mathbf{x_0} = [1 \quad 1 \quad 1]^\mathsf{T}$. Do 8 steps. Verify that $\mathbf{A}$ has the spectrum $\{0, 3, 5\}$.

12. CAS EXPERIMENT. Power Method with Scaling. Shifting. (a) Write a program for $n \times n$ matrices that prints every step. Apply it to the (nonsymmetric!) matrix (20 steps), starting from $[1 \quad 1 \quad 1]^\mathsf{T}$.

$$\mathbf{A} = \begin{bmatrix} 15 & 12 & 3 \\ 18 & 44 & 18 \\ -19 & -36 & -7 \end{bmatrix}.$$

(b) Experiment in (a) with shifting. Which shift do you find optimal?

(c) Write a program as in (a) but for symmetric matrices that prints vectors, scaled vectors, q, and δ. Apply it to the matrix in Prob. 8.

(d). Optimality of δ. Consider $\mathbf{A} = \begin{bmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{bmatrix}$ and take $\mathbf{x_0} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$. Show that $q = 0$, $\delta = 1$ for all steps and the eigenvalues are ± 1, so that the interval $[q - \delta, q + \delta]$ cannot be shortened (by omitting ± 1) without losing the inclusion property. Experiment with other $\mathbf{x_0}$'s.

(e) Find a (nonsymmetric) matrix for which δ in (2) is no longer an error bound.

(f) Experiment systematically with speed of convergence by choosing matrices with the second greatest eigenvalue (i) almost equal to the greatest, (ii) somewhat different, (iii) much different.

20.9 Tridiagonalization and QR-Factorization

We consider the problem of computing *all* the eigenvalues of a ***real symmetric*** matrix $\mathbf{A} = [a_{jk}]$, discussing a method widely used in practice. In the ***first stage*** we reduce the given matrix stepwise to a **tridiagonal matrix**, that is, a matrix having all its nonzero entries on the main diagonal and in the positions immediately adjacent to the main diagonal (such as $\mathbf{A_3}$ in Fig. 450, Third Step). This reduction was invented by A. S. Householder[11] (*J. Assn. Comput. Machinery* **5** (1958), 335–342). See also Ref. [E29] in App. 1.

This Householder tridiagonalization will simplify the matrix without changing its eigenvalues. The latter will then be determined (approximately) by factoring the tridiagonalized matrix, as discussed later in this section.

[11]ALSTON SCOTT HOUSEHOLDER (1904–1993), American mathematician, known for his work in numerical analysis and mathematical biology. He was head of the mathematics division at Oakridge National Laboratory and later professor at the University of Tennessee. He was both president of ACM (Association for Computing Machinery) 1954–1956 and SIAM (Society for Industrial and Applied Mathematics) 1963–1964.

www.EngineeringEBooksPdf.com

Householder's Tridiagonalization Method[11]

An $n \times n$ real symmetric matrix $\mathbf{A} = [a_{jk}]$ being given, we reduce it by $n - 2$ successive similarity transformations (see Sec. 20.6) involving matrices $\mathbf{P}_1, \cdots, \mathbf{P}_{n-2}$ to tridiagonal form. These matrices are orthogonal and symmetric. Thus $\mathbf{P}_1^{-1} = \mathbf{P}_1^{\mathsf{T}} = \mathbf{P}_1$ and similarly for the others. These transformations produce, from the given $\mathbf{A}_0 = \mathbf{A} = [a_{jk}]$, the matrices $\mathbf{A}_1 = [a_{jk}^{(1)}], \mathbf{A}_2 = [a_{jk}^{(2)}], \cdots, \mathbf{A}_{n-2} = [a_{jk}^{(n-2)}]$ in the form

$$
\begin{aligned}
\mathbf{A}_1 &= \mathbf{P}_1\mathbf{A}_0\mathbf{P}_1 \\
\mathbf{A}_2 &= \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2 \\
&\cdots\cdots\cdots\cdots \\
\mathbf{B} = \mathbf{A}_{n-2} &= \mathbf{P}_{n-2}\mathbf{A}_{n-3}\mathbf{P}_{n-2}.
\end{aligned}
\tag{1}
$$

The transformations (1) create the necessary zeros, in the first step in Row 1 and Column 1, in the second step in Row 2 and Column 2, etc., as Fig. 450 illustrates for a 5×5 matrix. $\mathbf{B}$ is tridiagonal.

$$
\begin{bmatrix} * & * & & & \\ * & * & * & * & * \\ & * & * & * & * \\ & * & * & * & * \\ & * & * & * & * \end{bmatrix}
\qquad
\begin{bmatrix} * & * & & & \\ * & * & * & & \\ & * & * & * & * \\ & & * & * & * \\ & & * & * & * \end{bmatrix}
\qquad
\begin{bmatrix} * & * & & & \\ * & * & * & & \\ & * & * & * & \\ & & * & * & * \\ & & & * & * \end{bmatrix}
$$

First Step	Second Step	Third Step
$\mathbf{A}_1 = \mathbf{P}_1\mathbf{A}\mathbf{P}_1$	$\mathbf{A}_2 = \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2$	$\mathbf{A}_3 = \mathbf{P}_3\mathbf{A}_2\mathbf{P}_3$

Fig. 450. Householder's method for a 5×5 matrix. Positions left blank are zeros created by the method.

How do we determine $\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_{n-2}$? Now, all these $\mathbf{P}_r$ are of the form

$$
\mathbf{P}_r = \mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} \qquad (r = 1, \cdots, n-2) \tag{2}
$$

where $\mathbf{I}$ is the $n \times n$ unit matrix and $\mathbf{v}_r = [v_{jr}]$ is a unit vector with its first r components 0; thus

$$
\mathbf{v}_1 = \begin{bmatrix} 0 \\ * \\ * \\ \vdots \\ * \end{bmatrix}, \qquad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ * \\ \vdots \\ * \end{bmatrix}, \qquad \cdots, \qquad \mathbf{v}_{n-2} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ * \\ * \end{bmatrix} \tag{3}
$$

where the asterisks denote the other components (which will be nonzero in general).

www.EngineeringEBooksPdf.com

Step 1. $\mathbf{v}_1$ has the components

$$
\begin{aligned}
&& v_{11} &= 0 \\
&\textbf{(a)} && v_{21} = \sqrt{\frac{1}{2}\left(1 + \frac{|a_{21}|}{S_1}\right)} \\
\textbf{(4)} &\textbf{(b)} && v_{j1} = \frac{a_{j1}\,\mathrm{sgn}\,a_{21}}{2v_{21}S_1} \qquad j = 3, 4, \cdots, n \\
&&& \text{where} \\
&\textbf{(c)} && S_1 = \sqrt{a_{21}^2 + a_{31}^2 + \cdots + a_{n1}^2}
\end{aligned}
$$

where $S_1 > 0$, and $\mathrm{sgn}\,a_{21} = +1$ if $a_{21} \geqq 0$ and $\mathrm{sgn}\,a_{21} = -1$ if $a_{21} < 0$. With this we compute $\mathbf{P}_1$ by (2) and then $\mathbf{A}_1$ by (1). This was the first step.

Step 2. We compute $\mathbf{v}_2$ by (4) with all subscripts increased by 1 and the a_{jk} replaced by $a_{jk}^{(1)}$, the entries of $\mathbf{A}_1$ just computed. Thus [see also (3)]

$$
\begin{aligned}
&v_{12} = v_{22} = 0 \\
(4^*) \qquad &v_{32} = \sqrt{\frac{1}{2}\left(1 + \frac{|a_{32}^{(1)}|}{S_2}\right)} \\
&v_{j2} = \frac{a_{j2}^{(1)}\,\mathrm{sgn}\,a_{32}^{(1)}}{2v_{32}S_2} \qquad j = 4, 5, \cdots, n
\end{aligned}
$$

where

$$S_2 = \sqrt{a_{32}^{(1)^2} + a_{42}^{(1)^2} + \cdots + a_{n2}^{(1)^2}}.$$

With this we compute $\mathbf{P}_2$ by (2) and then $\mathbf{A}_2$ by (1).

Step 3. We compute $\mathbf{v}_3$ by (4*) with all subscripts increased by 1 and the $a_{jk}^{(1)}$ replaced by the entries $a_{jk}^{(2)}$ of $\mathbf{A}_2$, and so on.

EXAMPLE 1 Householder Tridiagonalization

Tridiagonalize the real symmetric matrix

$$\mathbf{A} = \mathbf{A}_0 = \begin{bmatrix} 6 & 4 & 1 & 1 \\ 4 & 6 & 1 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 1 & 2 & 5 \end{bmatrix}.$$

Solution. ***Step 1.*** We compute $S_1^2 = 4^2 + 1^2 + 1^2 = 18$ from (4c). Since $a_{21} = 4 > 0$, we have $\mathrm{sgn}\,a_{21} = +1$ in (4b) and get from (4) by straightforward computation

www.EngineeringEBooksPdf.com

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ v_{21} \\ v_{31} \\ v_{41} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.98559856 \\ 0.11957316 \\ 0.11957316 \end{bmatrix}.$$

From this and (2),

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.94280904 & -0.23570227 & -0.23570227 \\ 0 & -0.23570227 & 0.97140452 & -0.02859548 \\ 0 & -0.23570227 & -0.02859548 & 0.97140452 \end{bmatrix}.$$

From the first line in (1) we now get

$$\mathbf{A}_1 = \mathbf{P}_1\mathbf{A}_0\mathbf{P}_1 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & -1 & -1 \\ 0 & -1 & \frac{9}{2} & \frac{3}{2} \\ 0 & -1 & \frac{3}{2} & \frac{9}{2} \end{bmatrix}.$$

Step 2. From (4*) we compute $S_2^2 = 2$ and

$$\mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ v_{32} \\ v_{42} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0.92387953 \\ 0.38268343 \end{bmatrix}.$$

From this and (2),

$$\mathbf{P}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 & -1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}.$$

The second line in (1) now gives

$$\mathbf{B}_2 = \mathbf{A}_2 = \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & \sqrt{2} & 0 \\ 0 & \sqrt{2} & 6 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

This matrix $\mathbf{B}$ is tridiagonal. Since our given matrix has order $n = 4$, we needed $n - 2 = 2$ steps to accomplish this reduction, as claimed. (Do you see that we got more zeros than we can expect in general?)

$\mathbf{B}$ is similar to $\mathbf{A}$, as we now show in general. This is essential because $\mathbf{B}$ thus has the same spectrum as $\mathbf{A}$, by Theorem 2 in Sec. 20.6. ■

B *Similar to* A. We assert that $\mathbf{B}$ in (1) is similar to $\mathbf{A} = \mathbf{A}_0$. The matrix $\mathbf{P}_r$ is symmetric; indeed,

$$\mathbf{P}_r^{\mathsf{T}} = (\mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^{\mathsf{T}} = \mathbf{I}^{\mathsf{T}} - 2(\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^{\mathsf{T}} = \mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} = \mathbf{P}_r$$

www.EngineeringEBooksPdf.com

Also, $\mathbf{P}_r$ is orthogonal because $\mathbf{v}_r$ is a unit vector, so that $\mathbf{v}_r^\mathsf{T}\mathbf{v}_r = 1$ and thus

$$\begin{aligned}\mathbf{P}_r\mathbf{P}_r^\mathsf{T} = \mathbf{P}_r^2 = (\mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^\mathsf{T})^2 &= \mathbf{I} - 4\mathbf{v}_r\mathbf{v}_r^\mathsf{T} + 4\mathbf{v}_r\mathbf{v}_r^\mathsf{T}\mathbf{v}_r\mathbf{v}_r^\mathsf{T} \\ &= \mathbf{I} - 4\mathbf{v}_r\mathbf{v}_r^\mathsf{T} + 4\mathbf{v}_r(\mathbf{v}_r^\mathsf{T}\mathbf{v}_r)\mathbf{v}_r^\mathsf{T} = \mathbf{I}.\end{aligned}$$

Hence $\mathbf{P}_r^{-1} = \mathbf{P}_r^\mathsf{T} = \mathbf{P}_r$ and from (1) we now obtain

$$\begin{aligned}\mathbf{B} &= \mathbf{P}_{n-2}\mathbf{A}_{n-3}\mathbf{P}_{n-2} = \cdots \\ \cdots &= \mathbf{P}_{n-2}\mathbf{P}_{n-3}\cdots\mathbf{P}_1\mathbf{A}\mathbf{P}_1\cdots\mathbf{P}_{n-3}\mathbf{P}_{n-2} \\ &= \mathbf{P}_{n-2}^{-1}\mathbf{P}_{n-3}^{-1}\cdots\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1\cdots\mathbf{P}_{n-3}\mathbf{P}_{n-2} \\ &= \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\end{aligned}$$

where $\mathbf{P} = \mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_{n-2}$. This proves our assertion. ■

QR-Factorization Method

In 1958 H. Rutishauser[12] of Switzerland proposed the idea of using the LU-factorization (Sec. 20.2; he called it LR-factorization) in solving eigenvalue problems. An improved version of Rutishauser's method (avoiding breakdown if certain submatrices become singular, etc.; see Ref. [E29]) is the QR-method, independently proposed by the American J. G. F. Francis (*Computer J.* **4** (1961–62), 265–271, 332–345) and the Russian V. N. Kublanovskaya (*Zhurnal Vych. Mat. i Mat. Fiz.* **1** (1961), 555–570). The QR-method uses the factorization **QR** with orthogonal **Q** and upper triangular **R.** We discuss the **QR**-method for a real ***symmetric*** matrix. (For extensions to general matrices see Ref. [E29] in App. 1.)

In this method we first transform a given real symmetric $n \times n$ matrix $\mathbf{A}$ into a tridiagonal matrix $\mathbf{B}_0 = \mathbf{B}$ by Householder's method. This creates many zeros and thus reduces the amount of further work. Then we compute $\mathbf{B}_1, \mathbf{B}_2, \cdots$ stepwise according to the following iteration method.

Step 1. Factor $\mathbf{B}_0 = \mathbf{Q}_0\mathbf{R}_0$ with orthogonal $\mathbf{Q}_0$ and upper triangular $\mathbf{R}_0$. Then compute $\mathbf{B}_1 = \mathbf{R}_0\mathbf{Q}_0$.

Step 2. Factor $\mathbf{B}_1 = \mathbf{Q}_1\mathbf{R}_1$. Then compute $\mathbf{B}_2 = \mathbf{R}_1\mathbf{Q}_1$.

General Step s + 1.

$$\textbf{(5)} \qquad \begin{aligned} &\textbf{(a)} \qquad \text{Factor } \mathbf{B}_s = \mathbf{Q}_s\mathbf{R}_s. \\ &\textbf{(b)} \qquad \text{Compute } \mathbf{B}_{s+1} = \mathbf{R}_s\mathbf{Q}_s. \end{aligned}$$

Here $\mathbf{Q}_s$ is orthogonal and $\mathbf{R}_s$ upper triangular. The factorization (5a) will be explained below.

$\mathbf{B}_{s+1}$ Similar to B. Convergence to a Diagonal Matrix. From (5a) we have $\mathbf{R}_s = \mathbf{Q}_s^{-1}\mathbf{B}_s$. Substitution into (5b) gives

$$\mathbf{B}_{s+1} = \mathbf{R}_s\mathbf{Q}_s = \mathbf{Q}_s^{-1}\mathbf{B}_s\mathbf{Q}_s. \tag{6}$$

[12]HEINZ RUTISHAUSER (1918–1970). Swiss mathematician, professor at ETH Zurich. Known for his pioneering work in numerics and computer science.

www.EngineeringEBooksPdf.com

Thus $\mathbf{B}_{s+1}$ is similar to $\mathbf{B}_s$. Hence $\mathbf{B}_{s+1}$ is similar to $\mathbf{B}_0 = \mathbf{B}$ for all s. By Theorem 2, Sec. 20.6, this implies that $\mathbf{B}_{s+1}$ has the same eigenvalues as $\mathbf{B}$.

Also, $\mathbf{B}_{s+1}$ is symmetric. This follows by induction. Indeed, $\mathbf{B}_0 = \mathbf{B}$ is symmetric. Assuming $\mathbf{B}_s$ to be symmetric, that is, $\mathbf{B}_s{}^\mathsf{T} = \mathbf{B}_s$, and using $\mathbf{Q}_s^{-1} = \mathbf{Q}_s{}^\mathsf{T}$ (since $\mathbf{Q}_s$ is orthogonal), we get from (6) the symmetry,

$$\mathbf{B}_{s+1}{}^\mathsf{T} = (\mathbf{Q}_s{}^\mathsf{T}\mathbf{B}_s\mathbf{Q}_s)^\mathsf{T} = Q_s{}^\mathsf{T}\mathbf{B}_s{}^\mathsf{T}\mathbf{Q}_s = \mathbf{Q}_s{}^\mathsf{T}\mathbf{B}_s\mathbf{Q}_s = \mathbf{B}_{s+1}.$$

If the eigenvalues of $\mathbf{B}$ are different in absolute value, say, $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, then

$$\lim_{s\to\infty} \mathbf{B}_s = \mathbf{D}$$

where $\mathbf{D}$ is diagonal, with main diagonal entries $\lambda_1, \lambda_2, \cdots, \lambda_n$. (Proof in Ref. [E29] listed in App. 1.)

How to Get the QR-Factorization, say, $\mathbf{B} = \mathbf{B}_0 = [b_{jk}] = \mathbf{Q}_0\mathbf{R}_0$. The tridiagonal matrix $\mathbf{B}$ has $n-1$ generally nonzero entries below the main diagonal. These are $b_{21}, b_{32}, \cdots, b_{n,n-1}$. We multiply $\mathbf{B}$ from the left by a matrix $\mathbf{C}_2$ such that $\mathbf{C}_2\mathbf{B} = [b_{jk}^{(2)}]$ has $b_{21}^{(2)} = 0$. We multiply this by a matrix $\mathbf{C}_3$ such that $\mathbf{C}_3\mathbf{C}_2\mathbf{B} = [b_{jk}^{(3)}]$ has $b_{32}^{(3)} = 0$, etc. After $n-1$ such multiplications we are left with an upper triangular matrix $\mathbf{R}_0$, namely,

$$\mathbf{C}_n\mathbf{C}_{n-1}\cdots\mathbf{C}_3\mathbf{C}_2B_0 = \mathbf{R}_0. \tag{7}$$

These $n \times n$ matrices $\mathbf{C}_j$ are very simple. $\mathbf{C}_j$ has the 2×2 submatrix

$$\begin{bmatrix} \cos\theta_j & \sin\theta_j \\ -\sin\theta_j & \cos\theta_j \end{bmatrix} \qquad (\theta_j \text{ suitable})$$

in Rows $j-1$ and j and Columns $j-1$ and j; everywhere else on the main diagonal the matrix $\mathbf{C}_j$ has entries 1; and all its other entries are 0. (This submatrix is the matrix of a plane rotation through the angle θ_j; see Team Project 30, Sec. 7.2.) For instance, if $n = 4$, writing $c_j = \cos\theta_j$, $s_j = \sin\theta_j$, we have

$$\mathbf{C}_2 = \begin{bmatrix} c_2 & s_2 & 0 & 0 \\ -s_2 & c_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{C}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_3 & s_3 & 0 \\ 0 & -s_3 & c_3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{C}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c_4 & s_4 \\ 0 & 0 & -s_4 & c_4 \end{bmatrix}.$$

These $\mathbf{C}_j$ are orthogonal. Hence their product in (7) is orthogonal, and so is the inverse of this product. We call this inverse $\mathbf{Q}_0$. Then from (7),

$$\mathbf{B}_0 = \mathbf{Q}_0\mathbf{R}_0 \tag{8}$$

where, with $\mathbf{C}_j^{-1} = \mathbf{C}_j{}^\mathsf{T}$,

$$\mathbf{Q}_0 = (\mathbf{C}_n\mathbf{C}_{n-1}\cdots\mathbf{C}_3\mathbf{C}_2)^{-1} = \mathbf{C}_2{}^\mathsf{T}\mathbf{C}_3{}^\mathsf{T}\cdots\mathbf{C}_{n-1}{}^\mathsf{T}\mathbf{C}_n{}^\mathsf{T}. \tag{9}$$

www.EngineeringEBooksPdf.com

This is our QR-factorization of $\mathbf{B}_0$. From it we have by (5b) with $s = 0$

$$\mathbf{B}_1 = \mathbf{R}_0\mathbf{Q}_0 = \mathbf{R}_0\mathbf{C}_2^{\mathsf{T}}\mathbf{C}_3^{\mathsf{T}}\cdots\mathbf{C}_{n-1}^{\mathsf{T}}\mathbf{C}_n^{\mathsf{T}}. \tag{10}$$

We do not need $\mathbf{Q}_0$ explicitly, but to get $\mathbf{B}_1$ from (10), we first compute $\mathbf{R}_0\mathbf{C}_2^{\mathsf{T}}$, then $(\mathbf{R}_0\mathbf{C}_2^{\mathsf{T}})\mathbf{C}_3^{\mathsf{T}}$, etc. Similarly in the further steps that produce $\mathbf{B}_2, \mathbf{B}_3, \cdots$.

Determination of $\cos\theta_j$ and $\sin\theta_j$. We finally show how to find the angles of rotation. $\cos\theta_2$ and $\sin\theta_2$ in $\mathbf{C}_2$ must be such that $b_{21}^{(2)} = 0$ in the product

$$\mathbf{C}_2\mathbf{B} = \begin{bmatrix} c_2 & s_2 & 0 & \cdots \\ -s_2 & c_2 & 0 & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix}\begin{bmatrix} b_{11} & b_{12} & b_{13} & \cdots \\ b_{21} & b_{22} & b_{23} & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

Now $b_{21}^{(2)}$ is obtained by multiplying the second row of $\mathbf{C}_2$ by the first column of $\mathbf{B}$,

$$b_{21}^{(2)} = -s_2b_{11} + c_2b_{21} = -(\sin\theta_2)b_{11} + (\cos\theta_2)b_{21} = 0.$$

Hence $\tan\theta_2 = s_2/c_2 = b_{21}/b_{11}$, and

$$\begin{aligned} \cos\theta_2 &= \frac{1}{\sqrt{1+\tan^2\theta_2}} = \frac{1}{\sqrt{1+(b_{21}/b_{11})^2}} \\ \sin\theta_2 &= \frac{\tan\theta_2}{\sqrt{1+\tan^2\theta_2}} = \frac{b_{21}/b_{11}}{\sqrt{1+(b_{21}/b_{11})^2}}. \end{aligned} \tag{11}$$

Similarly for $\theta_3, \theta_4, \cdots$. The next example illustrates all this.

EXAMPLE 2 QR-Factorization Method

Compute all the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} 6 & 4 & 1 & 1 \\ 4 & 6 & 1 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 1 & 2 & 5 \end{bmatrix}.$$

Solution. We first reduce $\mathbf{A}$ to tridiagonal form. Applying Householder's method, we obtain (see Example 1)

$$\mathbf{A}_2 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & \sqrt{2} & 0 \\ 0 & \sqrt{2} & 6 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

www.EngineeringEBooksPdf.com

From the characteristic determinant we see that $\mathbf{A}_2$, hence $\mathbf{A}$, has the eigenvalue 3. (Can you see this directly from $\mathbf{A}_2$?) Hence it suffices to apply the QR-method to the tridiagonal 3×3 matrix

$$\mathbf{B}_0 = \mathbf{B} = \begin{bmatrix} 6 & -\sqrt{18} & 0 \\ -\sqrt{18} & 7 & \sqrt{2} \\ 0 & \sqrt{2} & 6 \end{bmatrix}.$$

Step 1. We multiply $\mathbf{B}$ from the left by

$$\mathbf{C}_2 = \begin{bmatrix} \cos\theta_2 & \sin\theta_2 & 0 \\ -\sin\theta_2 & \cos\theta_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and then } \mathbf{C}_2\mathbf{B} \text{ by} \quad \mathbf{C}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_3 & \sin\theta_3 \\ 0 & -\sin\theta_3 & \cos\theta_3 \end{bmatrix}.$$

Here $(-\sin\theta_2) \cdot 6 + (\cos\theta_2)(-\sqrt{18}) = 0$ gives (11) $\cos\theta_2 = 0.81649658$ and $\sin\theta_2 = -0.57735027$. With these values we compute

$$\mathbf{C}_2\mathbf{B} = \begin{bmatrix} 7.34846923 & -7.50555350 & -0.81649658 \\ 0 & 3.26598632 & 1.15470054 \\ 0 & 1.41421356 & 6.00000000 \end{bmatrix}.$$

In $\mathbf{C}_3$ we get from $(-\sin\theta_3) \cdot 3.26598632 + (\cos\theta_3) \cdot 1.41421356 = 0$ the values $\cos\theta_3 = 0.91766294$ and $\sin\theta_3 = 0.39735971$. This gives

$$\mathbf{R}_0 = \mathbf{C}_3\mathbf{C}_2\mathbf{B} = \begin{bmatrix} 7.34846923 & -7.50555350 & -0.81649658 \\ 0 & 3.55902608 & 3.44378413 \\ 0 & 0 & 5.04714615 \end{bmatrix}.$$

From this we compute

$$\mathbf{B}_1 = \mathbf{R}_0\mathbf{C}_2{}^{\mathsf{T}}\mathbf{C}_3{}^{\mathsf{T}} = \begin{bmatrix} 10.33333333 & -2.05480467 & 0 \\ -2.05480467 & 4.03508772 & 2.00553251 \\ 0 & 2.00553251 & 4.63157895 \end{bmatrix}$$

which is symmetric and tridiagonal. The off-diagonal entries in $\mathbf{B}_1$ are still large in absolute value. Hence we have to go on.

Step 2. We do the same computations as in the first step, with $\mathbf{B}_0 = \mathbf{B}$ replaced by $\mathbf{B}_1$ and $\mathbf{C}_2$ and $\mathbf{C}_3$ changed accordingly, the new angles being $\theta_2 = -0.196291533$ and $\theta_3 = 0.513415589$. We obtain

$$\mathbf{R}_1 = \begin{bmatrix} 10.53565375 & -2.80232241 & -0.39114588 \\ 0 & 4.08329584 & 3.98824028 \\ 0 & 0 & 3.06832668 \end{bmatrix}$$

and from this

$$\mathbf{B}_2 = \begin{bmatrix} 10.87987988 & -0.79637918 & 0 \\ -0.79637918 & 5.44738664 & 1.50702500 \\ 0 & 1.50702500 & 2.67273348 \end{bmatrix}.$$

We see that the off-diagonal entries are somewhat smaller in absolute value than those of $\mathbf{B}_1$, but still much too large for the diagonal entries to be good approximations of the eigenvalues of $\mathbf{B}$.

www.EngineeringEBooksPdf.com

Further Steps. We list the main diagonal entries and the absolutely largest off-diagonal entry, which is $|b_{12}^{(j)}| = |b_{21}^{(j)}|$ in all steps. You may show that the given matrix **A** has the spectrum 11, 6, 3, 2.

Step j	$b_{11}^{(j)}$	$b_{22}^{(j)}$	$b_{33}^{(j)}$	$\max_{j \neq k} \lvert b_{jk}^{(J)} \rvert$
3	10.9668929	5.94589856	2.08720851	0.58523582
5	10.9970872	6.00181541	2.00109738	0.12065334
7	10.9997421	6.00024439	2.00001355	0.03591107
9	10.9999772	6.00002267	2.00000017	0.01068477

Looking back at our discussion, we recognize that the purpose of applying Householder's tridiagonalization before the QR-factorization method is a substantial reduction of cost in each QR-factorization, in particular if **A** is large.

Convergence acceleration and thus further reduction of cost can be achieved by a **spectral shift,** that is, by taking $\mathbf{B}_s - k_s\mathbf{I}$ instead of $\mathbf{B}_s$ with a suitable k_s. Possible choices of k_s are discussed in Ref. [E29], p. 510.

PROBLEM SET 20.9

1–5 HOUSEHOLDER TRIDIAGONALIZATION

Tridiagonalize. Show the details.

1. $\begin{bmatrix} 0.98 & 0.04 & 0.44 \\ 0.04 & 0.56 & 0.40 \\ 0.44 & 0.40 & 0.80 \end{bmatrix}$

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

3. $\begin{bmatrix} 7 & 2 & 3 \\ 2 & 10 & 6 \\ 3 & 6 & 7 \end{bmatrix}$

4. $\begin{bmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{bmatrix}$

5. $\begin{bmatrix} 3 & 52 & 10 & 42 \\ 52 & 59 & 44 & 80 \\ 10 & 44 & 39 & 42 \\ 42 & 80 & 42 & 35 \end{bmatrix}$

6–9 QR-FACTORIZATION

Do three QR-steps to find approximations of the eigenvalues of:

6. The matrix in the answer to Prob. 1

7. The matrix in the answer to Prob. 3

8. $\begin{bmatrix} 14.2 & -0.1 & 0 \\ -0.1 & -6.3 & 0.2 \\ 0 & 0.2 & 2.1 \end{bmatrix}$ **9.** $\begin{bmatrix} 140 & 10 & 0 \\ 10 & 70 & 2 \\ 0 & 2 & -30 \end{bmatrix}$

10. CAS EXPERIMENT. QR-Method. Try to find out experimentally on what properties of a matrix the speed of decrease of off-diagonal entries in the QR-method depends. For this purpose write a program that first tridiagonalizes and then does QR-steps. Try the program out on the matrices in Probs. 1, 3, and 4. Summarize your findings in a short report.

CHAPTER 20 REVIEW QUESTIONS AND PROBLEMS

1. What are the main problem areas in numeric linear algebra?

2. When would you apply Gauss elimination and when Gauss–Seidel iteration?

3. What is pivoting? Why and how is it done?

4. What happens if you apply Gauss elimination to a system that has no solutions?

5. What is Cholesky's method? When would you apply it?

www.EngineeringEBooksPdf.com

6. What do you know about the convergence of the Gauss–Seidel iteration?
7. What is ill-conditioning? What is the condition number and its significance?
8. Explain the idea of least squares approximation.
9. What are eigenvalues of a matrix? Why are they important? Give typical examples.
10. How did we use similarity transformations of matrices in designing numeric methods?
11. What is the power method for eigenvalues? What are its advantages and disadvantages?
12. State Gerschgorin's theorem from memory. Give typical applications.
13. What is tridiagonalization and QR? When would you apply it?

14–17 GAUSS ELIMINATION

Solve

14. $$\begin{aligned} 3x_2 - 6x_3 &= 0 \\ 4x_1 - x_2 + 2x_3 &= 16 \\ -5x_1 + 2x_2 - 4x_3 &= -20 \end{aligned}$$

15. $$\begin{aligned} 8x_2 - 6x_3 &= 23.6 \\ 10x_1 + 6x_2 + 2x_3 &= 68.4 \\ 12x_1 - 14x_2 + 4x_3 &= -6.2 \end{aligned}$$

16. $$\begin{aligned} 5x_1 + x_2 - 3x_3 &= 17 \\ -5x_2 + 15x_3 &= -10 \\ 2x_1 - 3x_2 + 9x_3 &= 0 \end{aligned}$$

17. $$\begin{aligned} 42x_1 + 74x_2 + 36x_3 &= 96 \\ -46x_1 - 12x_2 - 2x_3 &= 82 \\ 3x_1 + 25x_2 + 5x_3 &= 19 \end{aligned}$$

18–20 INVERSE MATRIX

Compute the inverse of:

18. $$\begin{bmatrix} 2.0 & 0.1 & 3.3 \\ 1.6 & 4.4 & 0.5 \\ 0.3 & -4.3 & 2.8 \end{bmatrix}$$

19. $$\begin{bmatrix} 15 & 20 & 10 \\ 20 & 35 & 15 \\ 10 & 15 & 90 \end{bmatrix}$$

20. $$\begin{bmatrix} 5 & 1 & 1 \\ 1 & 6 & 0 \\ 1 & 0 & 8 \end{bmatrix}$$

21–23 GAUSS–SEIDEL ITERATION

Do 3 steps without scaling, starting from $[1 \quad 1 \quad 1]^T$.

21. $$\begin{aligned} 4x_1 - x_2 \qquad &= 22.0 \\ 4x_2 - x_3 &= 13.4 \\ -x_1 \qquad + 4x_3 &= -2.4 \end{aligned}$$

22. $$\begin{aligned} 0.2x_1 + 4.0x_2 - 0.4x_3 &= 32.0 \\ 0.5x_1 - 0.2x_2 + 2.5x_3 &= -5.1 \\ 7.5x_1 + 0.1x_2 - 1.5x_3 &= -12.7 \end{aligned}$$

23. $$\begin{aligned} 10x_1 + x_2 - x_3 &= 17 \\ 2x_1 + 20x_2 + x_3 &= 28 \\ 3x_1 - x_2 + 25x_3 &= 105 \end{aligned}$$

24–26 VECTOR NORMS

Compute the ℓ_1-, ℓ_2-, and ℓ_∞-norms of the vectors.

24. $[0.2 \quad -8.1 \quad 0.4 \quad 0 \quad 0 \quad -1.3 \quad 2]^T$
25. $[8 \quad -21 \quad 13 \quad 0]^T$
26. $[0 \quad 0 \quad 0 \quad -1 \quad 0]^T$

27–30 MATRIX NORM

Compute the matrix norm corresponding to the ℓ_∞-vector norm for the coefficient matrix:

27. In Prob. 15
28. In Prob. 17
29. In Prob. 21
30. In Prob. 22

31–33 CONDITION NUMBER

Compute the condition number (corresponding to the ℓ_∞-vector norm) of the coefficient matrix:

31. In Prob. 19
32. In Prob. 18
33. In Prob. 21

34–35 FITTING BY LEAST SQUARES

Fit and graph:

34. A straight line to $(-1, 0)$, $(0, 2)$, $(1, 2)$, $(2, 3)$, $(3, 3)$
35. A quadratic parabola to the data in Prob. 34.

www.EngineeringEBooksPdf.com

36–39 EIGENVALUES

Find and graph three circular disks that must contain all the eigenvalues of the matrix:

36. In Prob. 18

37. In Prob. 19

38. In Prob. 20

39. Of the coefficients in Prob. 14

40. Power method. Do 4 steps with scaling for the matrix in Prob. 19, starting for [1 1 1] and computing the Rayliegh quotients and error bounds.

SUMMARY OF CHAPTER 20
Numeric Linear Algebra

Main tasks are the numeric solution of linear systems (Secs. 20.1–20.4), curve fitting (Sec. 20.5), and eigenvalue problems (Secs. 20.6–20.9).

Linear systems $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} = [a_{jk}]$, written out

$$
(1) \qquad \begin{aligned}
\mathrm{E}_1: &\quad a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\
\mathrm{E}_2: &\quad a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\mathrm{E}_n: &\quad a_{n1}x_1 + \cdots + a_{nn}x_n = b_n
\end{aligned}
$$

can be solved by a **direct method** (one in which the number of numeric operations can be specified in advance, e.g., Gauss's elimination) or by an **indirect** or **iterative method** (in which an initial approximation is improved stepwise).

The **Gauss elimination** (Sec. 20.1) is direct, namely, a systematic elimination process that reduces (1) stepwise to triangular form. In Step 1 we eliminate x_1 from equations E_2 to E_n by subtracting $(a_{21}/a_{11})\,\mathrm{E}_1$ from E_2, then $(a_{31}/a_{11})\,\mathrm{E}_1$ from E_3, etc. Equation E_1 is called the **pivot equation** in this step and a_{11} the **pivot.** In Step 2 we take the new second equation as pivot equation and eliminate x_2, etc. If the triangular form is reached, we get x_n from the last equation, then x_{n-1} from the second last, etc. **Partial pivoting** (= interchange of equations) is *necessary* if candidates for pivots are zero, and *advisable* if they are small in absolute value.

Doolittle's, Crout's, and **Cholesky's methods** in Sec. 20.2 are variants of the Gauss elimination. They factor $\mathbf{A} = \mathbf{LU}$ ($\mathbf{L}$ lower triangular, $\mathbf{U}$ upper triangular) and solve $\mathbf{Ax} = \mathbf{LUx} = \mathbf{b}$ by solving $\mathbf{Ly} = \mathbf{b}$ for $\mathbf{y}$ and then $\mathbf{Ux} = \mathbf{y}$ for $\mathbf{x}$.

In the **Gauss–Seidel iteration** (Sec. 20.3) we make $a_{11} = a_{22} = \cdots = a_{nn} = 1$ (by division) and write $\mathbf{Ax} = (\mathbf{I} + \mathbf{L} + \mathbf{U})\mathbf{x} = \mathbf{b}$; thus $\mathbf{x} = \mathbf{b} - (\mathbf{L} + \mathbf{U})\mathbf{x}$, which suggests the iteration formula

$$
(2) \qquad \mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - \mathbf{U}\mathbf{x}^{(m)}
$$

in which we always take the most recent approximate x_j's on the right. If $\|\mathbf{C}\| < 1$, where $\mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U}$, then this process converges. Here, $\|\mathbf{C}\|$ denotes any matrix norm (Sec. 20.3).

www.EngineeringEBooksPdf.com

If the **condition number** $k(\mathbf{A}) = \|\mathbf{A}\| \, \|\mathbf{A}^{-1}\|$ of $\mathbf{A}$ is large, then the system $\mathbf{Ax} = \mathbf{b}$ is **ill-conditioned** (Sec. 20.4), and a small **residual** $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ does ***not*** imply that $\tilde{\mathbf{x}}$ is close to the exact solution.

The fitting of a polynomial $p(x) = b_0 + b_1 x + \cdots + b_m x^m$ through given data (points in the xy-plane) $(x_1, y_1), \cdots, (x_n, y_n)$ by the method of **least squares** is discussed in Sec. 20.5 (and in statistics in Sec. 25.9). If $m = n$, the least squares polynomial will be the same as an interpolating polynomial (uniqueness).

Eigenvalues λ (values λ for which $\mathbf{Ax} = \lambda\mathbf{x}$ has a solution $\mathbf{x} \neq \mathbf{0}$, called an **eigenvector**) can be characterized by inequalities (Sec. 20.7), e.g. in **Gerschgorin's theorem,** which gives n circular disks which contain the whole spectrum (all eigenvalues) of $\mathbf{A}$, of centers a_{jj} and radii $\Sigma|a_{jk}|$ (sum over k from 1 to n, $k \neq j$).

Approximations of eigenvalues can be obtained by iteration, starting from an $\mathbf{x}_0 \neq \mathbf{0}$ and computing $\mathbf{x}_1 = \mathbf{Ax}_0$, $\mathbf{x}_2 = \mathbf{Ax}_1, \cdots, \mathbf{x}_n = \mathbf{Ax}_{n-1}$. In this **power method** (Sec. 20.8) the **Rayleigh quotient**

$$q = \frac{(\mathbf{Ax})^{\mathsf{T}})\mathbf{x}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} \qquad (\mathbf{x} = \mathbf{x}_n) \tag{3}$$

gives an approximation of an eigenvalue (usually that of the greatest absolute value) and, if $\mathbf{A}$ is symmetric, an error bound is

$$|\epsilon| \leqq \sqrt{\frac{(\mathbf{Ax})^{\mathsf{T}}\mathbf{Ax}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} - q^2}. \tag{4}$$

Convergence may be slow but can be improved by a *spectral shift.*

For determining all the eigenvalues of a symmetric matrix $\mathbf{A}$ it is best to first tridiagonalize $\mathbf{A}$ and then to apply the QR-method (Sec. 20.9), which is based on a factorization $\mathbf{A} = \mathbf{QR}$ with orthogonal $\mathbf{Q}$ and upper triangular $\mathbf{R}$ and uses similarity transformations.

www.EngineeringEBooksPdf.com

CHAPTER

2 Basic Structures: Sets, Functions, Sequences, and Sums

2.1 Sets

2.2 Set Operations

2.3 Functions

2.4 Sequences and Summations

Much of discrete mathematics is devoted to the study of discrete structures, used to represent discrete objects. Many important discrete structures are built using sets, which are collections of objects. Among the discrete structures built from sets are combinations, unordered collections of objects used extensively in counting; relations, sets of ordered pairs that represent relationships between objects; graphs, sets of vertices and edges that connect vertices; and finite state machines, used to model computing machines. These are some of the topics we will study in later chapters.

The concept of a function is extremely important in discrete mathematics. A function assigns to each element of a set exactly one element of a set. Functions play important roles throughout discrete mathematics. They are used to represent the computational complexity of algorithms, to study the size of sets, to count objects, and in a myriad of other ways. Useful structures such as sequences and strings are special types of functions. In this chapter, we will introduce the notion of a sequence, which represents ordered lists of elements. We will introduce some important types of sequences, and we will address the problem of identifying a pattern for the terms of a sequence from its first few terms. Using the notion of a sequence, we will define what it means for a set to be countable, namely, that we can list all the elements of the set in a sequence.

In our study of discrete mathematics, we will often add consecutive terms of a sequence of numbers. Because adding terms from a sequence, as well as other indexed sets of numbers, is such a common occurrence, a special notation has been developed for adding such terms. In this section, we will introduce the notation used to express summations. We will develop formulae for certain types of summations. Such summations appear throughout the study of discrete mathematics, as, for instance, when we analyze the number of steps a procedure uses to sort a list of numbers into increasing order.

2.1 Sets

Introduction

In this section, we study the fundamental discrete structure on which all other discrete structures are built, namely, the set. Sets are used to group objects together. Often, the objects in a set have similar properties. For instance, all the students who are currently enrolled in your school make up a set. Likewise, all the students currently taking a course in discrete mathematics at any school make up a set. In addition, those students enrolled in your school who are taking a course in discrete mathematics form a set that can be obtained by taking the elements common to the first two collections. The language of sets is a means to study such collections in an organized fashion. We now provide a definition of a set. This definition is an intuitive definition, which is not part of a formal theory of sets.

DEFINITION 1 A *set* is an unordered collection of objects.

Note that the term *object* has been used without specifying what an object is. This description of a set as a collection of objects, based on the intuitive notion of an object, was first stated by the German mathematician Georg Cantor in 1895. The theory that results from this intuitive

definition of a set, and the use of the intuitive notion that any property whatever there is a set consisting of exactly the objects with this property, leads to **paradoxes,** or logical inconsistencies. This was shown by the English philosopher Bertrand Russell in 1902 (see Exercise 38 for a description of one of these paradoxes). These logical inconsistencies can be avoided by building set theory beginning with axioms. We will use Cantor's original version of set theory, known as **naive set theory,** without developing an axiomatic version of set theory, because all sets considered in this book can be treated consistently using Cantor's original theory.

Assessment

DEFINITION 2 The objects in a set are called the *elements,* or *members,* of the set. A set is said to *contain* its elements.

We will now introduce notation used to describe membership in sets. We write $a \in A$ to denote that a is an element of the set A. The notation $a \notin A$ denotes that a is not an element of the set A. Note that lowercase letters are usually used to denote elements of sets.

There are several ways to describe a set. One way is to list all the members of a set, when this is possible. We use a notation where all members of the set are listed between braces. For example, the notation $\{a, b, c, d\}$ represents the set with the four elements a, b, c, and d.

EXAMPLE 1 The set V of all vowels in the English alphabet can be written as $V = \{a, e, i, o, u\}$. ◀

EXAMPLE 2 The set O of odd positive integers less than 10 can be expressed by $O = \{1, 3, 5, 7, 9\}$. ◀

EXAMPLE 3 Although sets are usually used to group together elements with common properties, there is nothing that prevents a set from having seemingly unrelated elements. For instance, $\{a$, 2, Fred, New Jersey$\}$ is the set containing the four elements a, 2, Fred, and New Jersey. ◀

Sometimes the brace notation is used to describe a set without listing all its members. Some members of the set are listed, and then *ellipses* (...) are used when the general pattern of the elements is obvious.

EXAMPLE 4 The set of positive integers less than 100 can be denoted by $\{1, 2, 3, \ldots, 99\}$. ◀

Extra Examples

Another way to describe a set is to use **set builder** notation. We characterize all those elements in the set by stating the property or properties they must have to be members. For instance, the set O of all odd positive integers less than 10 can be written as

$$O = \{x \mid x \text{ is an odd positive integer less than } 10\},$$

or, specifying the universe as the set of positive integers, as

$$O = \{x \in \mathbf{Z}^+ \mid x \text{ is odd and } x < 10\}.$$

We often use this type of notation to describe sets when it is impossible to list all the elements of the set. For instance, the set $\mathbf{Q}^+$ of all positive rational numbers can be written as

$$\mathbf{Q}^+ = \{x \in \mathbf{R} \mid x = p/q, \text{ for some positive integers } p \text{ and } q.$$

These sets, each denoted using a boldface letter, play an important role in discrete mathematics:

$\mathbf{N} = \{0, 1, 2, 3, \ldots\}$, the set of **natural numbers**
$\mathbf{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$, the set of **integers**

$\mathbf{Z}^+ = \{1, 2, 3, \ldots\}$, the set of **positive integers**
$\mathbf{Q} = \{p/q \mid p \in \mathbf{Z}, q \in \mathbf{Z}, \text{ and } q \neq 0\}$, the set of **rational numbers**
$\mathbf{R}$, the set of **real numbers**

(Note that some people do not consider 0 a natural number, so be careful to check how the term *natural numbers* is used when you read other books.)

Sets can have other sets as members, as Example 5 illustrates.

EXAMPLE 5 The set $\{\mathbf{N}, \mathbf{Z}, \mathbf{Q}, \mathbf{R}\}$ is a set containing four elements, each of which is a set. The four elements of this set are $\mathbf{N}$, the set of natural numbers; $\mathbf{Z}$, the set of integers; $\mathbf{Q}$, the set of rational numbers; and $\mathbf{R}$, the set of real numbers. ◀

Remark: Note that the concept of a datatype, or type, in computer science is built upon the concept of a set. In particular, a **datatype** or **type** is the name of a set, together with a set of operations that can be performed on objects from that set. For example, *boolean* is the name of the set $\{0, 1\}$ together with operators on one or more elements of this set, such as AND, OR, and NOT.

Because many mathematical statements assert that two differently specified collections of objects are really the same set, we need to understand what it means for two sets to be equal.

DEFINITION 3 Two sets are *equal* if and only if they have the same elements. That is, if A and B are sets, then A and B are equal if and only if $\forall x(x \in A \leftrightarrow x \in B)$. We write $A = B$ if A and B are equal sets.

EXAMPLE 6 The sets $\{1, 3, 5\}$ and $\{3, 5, 1\}$ are equal, because they have the same elements. Note that the order in which the elements of a set are listed does not matter. Note also that it does not matter if an element of a set is listed more than once, so $\{1, 3, 3, 3, 5, 5, 5, 5\}$ is the same as the set $\{1, 3, 5\}$ because they have the same elements. ◀

Sets can be represented graphically using Venn diagrams, named after the English mathematician John Venn, who introduced their use in 1881. In Venn diagrams the **universal set** U, which contains all the objects under consideration, is represented by a rectangle. (Note that the universal set varies depending on which objects are of interest.) Inside this rectangle, circles or other geometrical figures are used to represent sets. Sometimes points are used to represent the particular elements of the set. Venn diagrams are often used to indicate the relationships between sets. We show how a Venn diagram can be used in Example 7.

Links

GEORG CANTOR (1845–1918) Georg Cantor was born in St. Petersburg, Russia, where his father was a successful merchant. Cantor developed his interest in mathematics in his teens. He began his university studies in Zurich in 1862, but when his father died he left Zurich. He continued his university studies at the University of Berlin in 1863, where he studied under the eminent mathematicians Weierstrass, Kummer, and Kronecker. He received his doctor's degree in 1867, after having written a dissertation on number theory. Cantor assumed a position at the University of Halle in 1869, where he continued working until his death.

Cantor is considered the founder of set theory. His contributions in this area include the discovery that the set of real numbers is uncountable. He is also noted for his many important contributions to analysis. Cantor also was interested in philosophy and wrote papers relating his theory of sets with metaphysics.

Cantor married in 1874 and had five children. His melancholy temperament was balanced by his wife's happy disposition. Although he received a large inheritance from his father, he was poorly paid as a professor. To mitigate this, he tried to obtain a better-paying position at the University of Berlin. His appointment there was blocked by Kronecker, who did not agree with Cantor's views on set theory. Cantor suffered from mental illness throughout the later years of his life. He died in 1918 in a psychiatric clinic.

FIGURE 1 Venn Diagram for the Set of Vowels.

EXAMPLE 7 Draw a Venn diagram that represents V, the set of vowels in the English alphabet.

Solution: We draw a rectangle to indicate the universal set U, which is the set of the 26 letters of the English alphabet. Inside this rectangle we draw a circle to represent V. Inside this circle we indicate the elements of V with points (see Figure 1). ◀

There is a special set that has no elements. This set is called the **empty set,** or **null set,** and is denoted by $\emptyset$. The empty set can also be denoted by { } (that is, we represent the empty set with a pair of braces that encloses all the elements in this set). Often, a set of elements with certain properties turns out to be the null set. For instance, the set of all positive integers that are greater than their squares is the null set. A set with one element is called a **singleton set.**

A common error is to confuse the empty set $\emptyset$ with the set $\{\emptyset\}$, which is a singleton set. The single element of the set $\{\emptyset\}$ is the empty set itself! A useful analogy for remembering this difference is to think of folders in a computer file system. The empty set can be thought of as an empty folder and the set consisting of just the empty set can be thought of as a folder with exactly one folder inside, namely, the empty folder.

DEFINITION 4 The set A is said to be a *subset* of B if and only if every element of A is also an element of B. We use the notation $A \subseteq B$ to indicate that A is a subset of the set B.

We see that $A \subseteq B$ if and only if the quantification

$$\forall x(x \in A \rightarrow x \in B)$$

is true.

Links

BERTRAND RUSSELL (1872–1970) Bertrand Russell was born into a prominent English family active in the progressive movement and having a strong commitment to liberty. He became an orphan at an early age and was placed in the care of his father's parents, who had him educated at home. He entered Trinity College, Cambridge, in 1890, where he excelled in mathematics and in moral science. He won a fellowship on the basis of his work on the foundations of geometry. In 1910 Trinity College appointed him to a lectureship in logic and the philosophy of mathematics.

Russell fought for progressive causes throughout his life. He held strong pacifist views, and his protests against World War I led to dismissal from his position at Trinity College. He was imprisoned for 6 months in 1918 because of an article he wrote that was branded as seditious. Russell fought for women's suffrage in Great Britain. In 1961, at the age of 89, he was imprisoned for the second time for his protests advocating nuclear disarmament.

Russell's greatest work was in his development of principles that could be used as a foundation for all of mathematics. His most famous work is *Principia Mathematica,* written with Alfred North Whitehead, which attempts to deduce all of mathematics using a set of primitive axioms. He wrote many books on philosophy, physics, and his political ideas. Russell won the Nobel Prize for literature in 1950.

FIGURE 2 Venn Diagram Showing that A Is a Subset of B.

EXAMPLE 8 The set of all odd positive integers less than 10 is a subset of the set of all positive integers less than 10, the set of rational numbers is a subset of the set of real numbers, the set of all computer science majors at your school is a subset of the set of all students at your school, and the set of all people in China is a subset of the set of all people in China (that is, it is a subset of itself). ◀

Theorem 1 shows that every nonempty set S is guaranteed to have at least two subsets, the empty set and the set S itself, that is, $\emptyset \subseteq S$ and $S \subseteq S$.

THEOREM 1 For every set S,
$(i)\ \emptyset \subseteq S$ and $(ii)\ S \subseteq S$.

Proof: We will prove (i) and leave the proof of (ii) as an exercise.

Let S be a set. To show that $\emptyset \subseteq S$, we must show that $\forall x(x \in \emptyset \rightarrow x \in S)$ is true. Because the empty set contains no elements, it follows that $x \in \emptyset$ is always false. It follows that the conditional statement $x \in \emptyset \rightarrow x \in S$ is always true, because its hypothesis is always false and a conditional statement with a false hypothesis is true. That is, $\forall x(x \in \emptyset \rightarrow x \in S)$ is true. This completes the proof of (i). Note that this is an example of a vacuous proof. ◁

When we wish to emphasize that a set A is a subset of the set B but that $A \neq B$, we write $A \subset B$ and say that A is a **proper subset** of B. For $A \subset B$ to be true, it must be the case that $A \subseteq B$ and there must exist an element x of B that is not an element of A. That is, A is a proper subset of B if

$$\forall x(x \in A \rightarrow x \in B) \wedge \exists x(x \in B \wedge x \notin A)$$

is true. Venn diagrams can be used to illustrate that a set A is a subset of a set B. We draw the universal set U as a rectangle. Within this rectangle we draw a circle for B. Because A is a subset of B, we draw the circle for A within the circle for B. This relationship is shown in Figure 2.

Links

JOHN VENN (1834–1923) John Venn was born into a London suburban family noted for its philanthropy. He attended London schools and got his mathematics degree from Caius College, Cambridge, in 1857. He was elected a fellow of this college and held his fellowship there until his death. He took holy orders in 1859 and, after a brief stint of religious work, returned to Cambridge, where he developed programs in the moral sciences. Besides his mathematical work, Venn had an interest in history and wrote extensively about his college and family.

Venn's book *Symbolic Logic* clarifies ideas originally presented by Boole. In this book, Venn presents a systematic development of a method that uses geometric figures, known now as *Venn diagrams*. Today these diagrams are primarily used to analyze logical arguments and to illustrate relationships between sets. In addition to his work on symbolic logic, Venn made contributions to probability theory described in his widely used textbook on that subject.

One way to show that two sets have the same elements is to show that each set is a subset of the other. In other words, we can show that if A and B are sets with $A \subseteq B$ and $B \subseteq A$, then $A = B$. This turns out to be a useful way to show that two sets are equal. That is, $A = B$, where A and B are sets, if and only if $\forall x(x \in A \rightarrow x \in B)$ and $\forall x(x \in B \rightarrow x \in A)$, or equivalently if and only if $\forall x(x \in A \leftrightarrow x \in B)$.

Sets may have other sets as members. For instance, we have the sets

$$A = \{\emptyset, \{a\}, \{b\}, \{a, b\}\} \qquad \text{and} \qquad B = \{x \mid x \text{ is a subset of the set } \{a, b\}\}.$$

Note that these two sets are equal, that is, $A = B$. Also note that $\{a\} \in A$, but $a \notin A$.

Sets are used extensively in counting problems, and for such applications we need to discuss the size of sets.

DEFINITION 5 Let S be a set. If there are exactly n distinct elements in S where n is a nonnegative integer, we say that S is a *finite set* and that n is the *cardinality* of S. The cardinality of S is denoted by $|S|$.

EXAMPLE 9 Let A be the set of odd positive integers less than 10. Then $|A| = 5$. ◀

EXAMPLE 10 Let S be the set of letters in the English alphabet. Then $|S| = 26$. ◀

EXAMPLE 11 Because the null set has no elements, it follows that $|\emptyset| = 0$. ◀

We will also be interested in sets that are not finite.

DEFINITION 6 A set is said to be *infinite* if it is not finite.

EXAMPLE 12 The set of positive integers is infinite. ◀

Extra Examples

The cardinality of infinite sets will be discussed in Section 2.4.

The Power Set

Many problems involve testing all combinations of elements of a set to see if they satisfy some property. To consider all such combinations of elements of a set S, we build a new set that has as its members all the subsets of S.

DEFINITION 7 Given a set S, the *power set* of S is the set of all subsets of the set S. The power set of S is denoted by $P(S)$.

EXAMPLE 13 What is the power set of the set $\{0, 1, 2\}$?

Solution: The power set $P(\{0, 1, 2\})$ is the set of all subsets of $\{0, 1, 2\}$. Hence,

$$P(\{0, 1, 2\}) = \{\emptyset, \{0\}, \{1\}, \{2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \{0, 1, 2\}\}.$$

Note that the empty set and the set itself are members of this set of subsets. ◀

EXAMPLE 14 What is the power set of the empty set? What is the power set of the set $\{\emptyset\}$?

Solution: The empty set has exactly one subset, namely, itself. Consequently,

$$P(\emptyset) = \{\emptyset\}.$$

The set $\{\emptyset\}$ has exactly two subsets, namely, $\emptyset$ and the set $\{\emptyset\}$ itself. Therefore,

$$P(\{\emptyset\}) = \{\emptyset, \{\emptyset\}\}.$$ ◀

If a set has n elements, then its power set has 2^n elements. We will demonstrate this fact in several ways in subsequent sections of the text.

Cartesian Products

The order of elements in a collection is often important. Because sets are unordered, a different structure is needed to represent ordered collections. This is provided by **ordered n-tuples.**

DEFINITION 8 The *ordered n-tuple* $(a_1, a_2, \ldots, a_n)$ is the ordered collection that has a_1 as its first element, a_2 as its second element, ..., and a_n as its nth element.

We say that two ordered n-tuples are equal if and only if each corresponding pair of their elements is equal. In other words, $(a_1, a_2, \ldots, a_n) = (b_1, b_2, \ldots, b_n)$ if and only if $a_i = b_i$, for $i = 1, 2, \ldots, n$. In particular, 2-tuples are called **ordered pairs.** The ordered pairs (a, b) and (c, d) are equal if and only if $a = c$ and $b = d$. Note that (a, b) and (b, a) are not equal unless $a = b$.

Many of the discrete structures we will study in later chapters are based on the notion of the *Cartesian product* of sets (named after René Descartes). We first define the Cartesian product of two sets.

Links

RENÉ DESCARTES (1596–1650) René Descartes was born into a noble family near Tours, France, about 200 miles southwest of Paris. He was the third child of his father's first wife; she died several days after his birth. Because of René's poor health, his father, a provincial judge, let his son's formal lessons slide until, at the age of 8, René entered the Jesuit college at La Flèche. The rector of the school took a liking to him and permitted him to stay in bed until late in the morning because of his frail health. From then on, Descartes spent his mornings in bed; he considered these times his most productive hours for thinking.

Descartes left school in 1612, moving to Paris, where he spent 2 years studying mathematics. He earned a law degree in 1616 from the University of Poitiers. At 18 Descartes became disgusted with studying and decided to see the world. He moved to Paris and became a successful gambler. However, he grew tired of bawdy living and moved to the suburb of Saint-Germain, where he devoted himself to mathematical study. When his gambling friends found him, he decided to leave France and undertake a military career. However, he never did any fighting. One day, while escaping the cold in an overheated room at a military encampment, he had several feverish dreams, which revealed his future career as a mathematician and philosopher.

After ending his military career, he traveled throughout Europe. He then spent several years in Paris, where he studied mathematics and philosophy and constructed optical instruments. Descartes decided to move to Holland, where he spent 20 years wandering around the country, accomplishing his most important work. During this time he wrote several books, including the *Discours,* which contains his contributions to analytic geometry, for which he is best known. He also made fundamental contributions to philosophy.

In 1649 Descartes was invited by Queen Christina to visit her court in Sweden to tutor her in philosophy. Although he was reluctant to live in what he called "the land of bears amongst rocks and ice," he finally accepted the invitation and moved to Sweden. Unfortunately, the winter of 1649–1650 was extremely bitter. Descartes caught pneumonia and died in mid-February.

DEFINITION 9 Let A and B be sets. The *Cartesian product* of A and B, denoted by $A \times B$, is the set of all ordered pairs (a, b), where $a \in A$ and $b \in B$. Hence,

$$A \times B = \{(a, b) \mid a \in A \wedge b \in B\}.$$

EXAMPLE 15 Let A represent the set of all students at a university, and let B represent the set of all courses offered at the university. What is the Cartesian product $A \times B$?

Extra Examples

Solution: The Cartesian product $A \times B$ consists of all the ordered pairs of the form (a, b), where a is a student at the university and b is a course offered at the university. The set $A \times B$ can be used to represent all possible enrollments of students in courses at the university. ◀

EXAMPLE 16 What is the Cartesian product of $A = \{1, 2\}$ and $B = \{a, b, c\}$?

Solution: The Cartesian product $A \times B$ is

$$A \times B = \{(1, a), (1, b), (1, c), (2, a), (2, b), (2, c)\}.$$ ◀

A subset R of the Cartesian product $A \times B$ is called a **relation** from the set A to the set B. The elements of R are ordered pairs, where the first element belongs to A and the second to B. For example, $R = \{(a, 0), (a, 1), (a, 3), (b, 1), (b, 2), (c, 0), (c, 3)\}$ is a relation from the set $\{a, b, c\}$ to the set $\{0, 1, 2, 3\}$. We will study relations at length in Chapter 8.

The Cartesian products $A \times B$ and $B \times A$ are not equal, unless $A = \emptyset$ or $B = \emptyset$ (so that $A \times B = \emptyset$) or $A = B$ (see Exercises 26 and 30, at the end of this section). This is illustrated in Example 17.

EXAMPLE 17 Show that the Cartesian product $B \times A$ is not equal to the Cartesian product $A \times B$, where A and B are as in Example 16.

Solution: The Cartesian product $B \times A$ is

$$B \times A = \{(a, 1), (a, 2), (b, 1), (b, 2), (c, 1), (c, 2)\}.$$

This is not equal to $A \times B$, which was found in Example 16. ◀

The Cartesian product of more than two sets can also be defined.

DEFINITION 10 The *Cartesian product* of the sets $A_1, A_2, \ldots, A_n$, denoted by $A_1 \times A_2 \times \cdots \times A_n$, is the set of ordered n-tuples $(a_1, a_2, \ldots, a_n)$, where a_i belongs to A_i for $i = 1, 2, \ldots, n$. In other words,

$$A_1 \times A_2 \times \cdots \times A_n = \{(a_1, a_2, \ldots, a_n) \mid a_i \in A_i \text{ for } i = 1, 2, \ldots, n\}.$$

EXAMPLE 18 What is the Cartesian product $A \times B \times C$, where $A = \{0, 1\}$, $B = \{1, 2\}$, and $C = \{0, 1, 2\}$?

Solution: The Cartesian product $A \times B \times C$ consists of all ordered triples (a, b, c), where $a \in A$, $b \in B$, and $c \in C$. Hence,

$$A \times B \times C = \{(0, 1, 0), (0, 1, 1), (0, 1, 2), (0, 2, 0), (0, 2, 1), (0, 2, 2), (1, 1, 0), (1, 1, 1), (1, 1, 2), (1, 2, 0), (1, 2, 1), (1, 2, 2)\}.$$ ◀

Using Set Notation with Quantifiers

Sometimes we restrict the domain of a quantified statement explicitly by making use of a particular notation. For example, $\forall x \in S(P(x))$ denotes the universal quantification of $P(x)$ over all elements in the set S. In other words, $\forall x \in S(P(x))$ is shorthand for $\forall x(x \in S \rightarrow P(x))$. Similarly, $\exists x \in S(P(x))$ denotes the existential quantification of $P(x)$ over all elements in S. That is, $\exists x \in S(P(x))$ is shorthand for $\exists x(x \in S \wedge P(x))$.

EXAMPLE 19 What do the statements $\forall x \in \mathbf{R}\ (x^2 \geq 0)$ and $\exists x \in \mathbf{Z}\ (x^2 = 1)$ mean?

Solution: The statement $\forall x \in \mathbf{R}(x^2 \geq 0)$ states that for every real number x, $x^2 \geq 0$. This statement can be expressed as "The square of every real number is nonnegative." This is a true statement.

The statement $\exists x \in \mathbf{Z}(x^2 = 1)$ states that there exists an integer x such that $x^2 = 1$. This statement can be expressed as "There is an integer whose square is 1." This is also a true statement because $x = 1$ is such an integer (as is -1). ◀

Truth Sets of Quantifiers

We will now tie together concepts from set theory and from predicate logic. Given a predicate P, and a domain D, we define the **truth set** of P to be the set of elements x in D for which $P(x)$ is true. The truth set of $P(x)$ is denoted by $\{x \in D \mid P(x)\}$.

EXAMPLE 20 What are the truth sets of the predicates $P(x)$, $Q(x)$, and $R(x)$, where the domain is the set of integers and $P(x)$ is "$|x| = 1$," $Q(x)$ is "$x^2 = 2$," and $R(x)$ is "$|x| = x$."

Solution: The truth set of P, $\{x \in \mathbf{Z} \mid |x| = 1\}$, is the set of integers for which $|x| = 1$. Because $|x| = 1$ when $x = 1$ or $x = -1$, and for no other integers x, we see that the truth set of P is the set $\{-1, 1\}$.

The truth set of Q, $\{x \in \mathbf{Z} \mid x^2 = 2\}$, is the set of integers for which $x^2 = 2$. This is the empty set because there are no integers x for which $x^2 = 2$.

The truth set of R, $\{x \in \mathbf{Z} \mid |x| = x\}$, is the set of integers for which $|x| = x$. Because $|x| = x$ if and only if $x \geq 0$, it follows that the truth set of R is $\mathbf{N}$, the set of nonnegative integers. ◀

Note that $\forall x P(x)$ is true over the domain U if and only if the truth set of P is the set U. Likewise, $\exists x P(x)$ is true over the domain U if and only if the truth set of P is nonempty.

Exercises

1. List the members of these sets.
   a) $\{x \mid x$ is a real number such that $x^2 = 1\}$
   b) $\{x \mid x$ is a positive integer less than $12\}$
   c) $\{x \mid x$ is the square of an integer and $x < 100\}$
   d) $\{x \mid x$ is an integer such that $x^2 = 2\}$
2. Use set builder notation to give a description of each of these sets.
   a) $\{0, 3, 6, 9, 12\}$
   b) $\{-3, -2, -1, 0, 1, 2, 3\}$
   c) $\{m, n, o, p\}$
3. Determine whether each of these pairs of sets are equal.
   a) $\{1, 3, 3, 3, 5, 5, 5, 5, 5\}, \{5, 3, 1\}$
   b) $\{\{1\}\}, \{1, \{1\}\}$ c) $\emptyset, \{\emptyset\}$
4. Suppose that $A = \{2, 4, 6\}$, $B = \{2, 6\}$, $C = \{4, 6\}$, and $D = \{4, 6, 8\}$. Determine which of these sets are subsets of which other of these sets.
5. For each of the following sets, determine whether 2 is an element of that set.
   a) $\{x \in \mathbf{R} \mid x$ is an integer greater than $1\}$
   b) $\{x \in \mathbf{R} \mid x$ is the square of an integer$\}$
   c) $\{2, \{2\}\}$ d) $\{\{2\}, \{\{2\}\}\}$
   e) $\{\{2\}, \{2, \{2\}\}\}$ f) $\{\{\{2\}\}\}$

6. For each of the sets in Exercise 5, determine whether $\{2\}$ is an element of that set.

7. Determine whether each of these statements is true or false.

a) $0 \in \emptyset$ b) $\emptyset \in \{0\}$
c) $\{0\} \subset \emptyset$ d) $\emptyset \subset \{0\}$
e) $\{0\} \in \{0\}$ f) $\{0\} \subset \{0\}$
g) $\{\emptyset\} \subseteq \{\emptyset\}$

8. Determine whether these statements are true or false.

a) $\emptyset \in \{\emptyset\}$ b) $\emptyset \in \{\emptyset, \{\emptyset\}\}$
c) $\{\emptyset\} \in \{\emptyset\}$ d) $\{\emptyset\} \in \{\{\emptyset\}\}$
e) $\{\emptyset\} \subset \{\emptyset, \{\emptyset\}\}$ f) $\{\{\emptyset\}\} \subset \{\emptyset, \{\emptyset\}\}$
g) $\{\{\emptyset\}\} \subset \{\{\emptyset\}, \{\emptyset\}\}$

9. Determine whether each of these statements is true or false.

a) $x \in \{x\}$ b) $\{x\} \subseteq \{x\}$ c) $\{x\} \in \{x\}$
d) $\{x\} \in \{\{x\}\}$ e) $\emptyset \subseteq \{x\}$ f) $\emptyset \in \{x\}$

10. Use a Venn diagram to illustrate the subset of odd integers in the set of all positive integers not exceeding 10.

11. Use a Venn diagram to illustrate the set of all months of the year whose names do not contain the letter R in the set of all months of the year.

12. Use a Venn diagram to illustrate the relationship $A \subseteq B$ and $B \subseteq C$.

13. Use a Venn diagram to illustrate the relationships $A \subset B$ and $B \subset C$.

14. Use a Venn diagram to illustrate the relationships $A \subset B$ and $A \subset C$.

15. Suppose that A, B, and C are sets such that $A \subseteq B$ and $B \subseteq C$. Show that $A \subseteq C$.

16. Find two sets A and B such that $A \in B$ and $A \subseteq B$.

17. What is the cardinality of each of these sets?

a) $\{a\}$ b) $\{\{a\}\}$
c) $\{a, \{a\}\}$ d) $\{a, \{a\}, \{a, \{a\}\}\}$

18. What is the cardinality of each of these sets?

a) $\emptyset$ b) $\{\emptyset\}$
c) $\{\emptyset, \{\emptyset\}\}$ d) $\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$

19. Find the power set of each of these sets, where a and b are distinct elements.

a) $\{a\}$ b) $\{a, b\}$ c) $\{\emptyset, \{\emptyset\}\}$

20. Can you conclude that $A = B$ if A and B are two sets with the same power set?

21. How many elements does each of these sets have where a and b are distinct elements?

a) $P(\{a, b, \{a, b\}\})$
b) $P(\{\emptyset, a, \{a\}, \{\{a\}\}\})$
c) $P(P(\emptyset))$

22. Determine whether each of these sets is the power set of a set, where a and b are distinct elements.

a) $\emptyset$ b) $\{\emptyset, \{a\}\}$
c) $\{\emptyset, \{a\}, \{\emptyset, a\}\}$ d) $\{\emptyset, \{a\}, \{b\}, \{a, b\}\}$

23. Let $A = \{a, b, c, d\}$ and $B = \{y, z\}$. Find

a) $A \times B$ b) $B \times A$

24. What is the Cartesian product $A \times B$, where A is the set of courses offered by the mathematics department at a university and B is the set of mathematics professors at this university?

25. What is the Cartesian product $A \times B \times C$, where A is the set of all airlines and B and C are both the set of all cities in the United States?

26. Suppose that $A \times B = \emptyset$, where A and B are sets. What can you conclude?

27. Let A be a set. Show that $\emptyset \times A = A \times \emptyset = \emptyset$.

28. Let $A = \{a, b, c\}$, $B = \{x, y\}$, and $C = \{0, 1\}$. Find

a) $A \times B \times C$ b) $C \times B \times A$
c) $C \times A \times B$ d) $B \times B \times B$

29. How many different elements does $A \times B$ have if A has m elements and B has n elements?

30. Show that $A \times B \neq B \times A$, when A and B are nonempty, unless $A = B$.

31. Explain why $A \times B \times C$ and $(A \times B) \times C$ are not the same.

32. Explain why $(A \times B) \times (C \times D)$ and $A \times (B \times C) \times D$ are not the same.

33. Translate each of these quantifications into English and determine its truth value.

a) $\forall x \in \mathbf{R}\,(x^2 \neq -1)$ b) $\exists x \in \mathbf{Z}\,(x^2 = 2)$
c) $\forall x \in \mathbf{Z}\,(x^2 > 0)$ d) $\exists x \in \mathbf{R}\,(x^2 = x)$

34. Translate each of these quantifications into English and determine its truth value.

a) $\exists x \in \mathbf{R}\,(x^3 = -1)$ b) $\exists x \in \mathbf{Z}\,(x + 1 > x)$
c) $\forall x \in \mathbf{Z}\,(x - 1 \in \mathbf{Z})$ d) $\forall x \in \mathbf{Z}\,(x^2 \in \mathbf{Z})$

35. Find the truth set of each of these predicates where the domain is the set of integers.

a) $P(x)$: "$x^2 < 3$" b) $Q(x)$: "$x^2 > x$"
c) $R(x)$: "$2x + 1 = 0$"

36. Find the truth set of each of these predicates where the domain is the set of integers.

a) $P(x)$: "$x^3 \geq 1$"
b) $Q(x)$: "$x^2 = 2$"
c) $R(x)$: "$x < x^2$"

*37. The defining property of an ordered pair is that two ordered pairs are equal if and only if their first elements are equal and their second elements are equal. Surprisingly, instead of taking the ordered pair as a primitive concept, we can construct ordered pairs using basic notions from set theory. Show that if we define the ordered pair (a, b) to be $\{\{a\}, \{a, b\}\}$, then $(a, b) = (c, d)$ if and only

if $a = c$ and $b = d$. [*Hint:* First show that $\{\{a\}, \{a, b\}\} = \{\{c\}, \{c, d\}\}$ if and only if $a = c$ and $b = d$.]

***38.** This exercise presents **Russell's paradox**. Let S be the set that contains a set x if the set x does not belong to itself, so that $S = \{x \mid x \notin x\}$.

a) Show the assumption that S is a member of S leads to a contradiction.

b) Show the assumption that S is not a member of S leads to a contradiction.

By parts (a) and (b) it follows that the set S cannot be defined as it was. This paradox can be avoided by restricting the types of elements that sets can have.

***39.** Describe a procedure for listing all the subsets of a finite set.

2.2 Set Operations

Introduction

Two sets can be combined in many different ways. For instance, starting with the set of mathematics majors at your school and the set of computer science majors at your school, we can form the set of students who are mathematics majors or computer science majors, the set of students who are joint majors in mathematics and computer science, the set of all students not majoring in mathematics, and so on.

Links

DEFINITION 1 Let A and B be sets. The *union* of the sets A and B, denoted by $A \cup B$, is the set that contains those elements that are either in A or in B, or in both.

An element x belongs to the union of the sets A and B if and only if x belongs to A or x belongs to B. This tells us that

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

The Venn diagram shown in Figure 1 represents the union of two sets A and B. The area that represents $A \cup B$ is the shaded area within either the circle representing A or the circle representing B.

We will give some examples of the union of sets.

EXAMPLE 1 The union of the sets $\{1, 3, 5\}$ and $\{1, 2, 3\}$ is the set $\{1, 2, 3, 5\}$; that is, $\{1, 3, 5\} \cup \{1, 2, 3\} = \{1, 2, 3, 5\}$. ◀

EXAMPLE 2 The union of the set of all computer science majors at your school and the set of all mathematics majors at your school is the set of students at your school who are majoring either in mathematics or in computer science (or in both). ◀

DEFINITION 2 Let A and B be sets. The *intersection* of the sets A and B, denoted by $A \cap B$, is the set containing those elements in both A and B.

An element x belongs to the intersection of the sets A and B if and only if x belongs to A and x belongs to B. This tells us that

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

$A \cup B$ is shaded.

FIGURE 1 Venn Diagram Representing the Union of A and B.

$A \cap B$ is shaded.

FIGURE 2 Venn Diagram Representing the Intersection of A and B.

The Venn diagram shown in Figure 2 represents the intersection of two sets A and B. The shaded area that is within both the circles representing the sets A and B is the area that represents the intersection of A and B.

We give some examples of the intersection of sets.

EXAMPLE 3 The intersection of the sets $\{1, 3, 5\}$ and $\{1, 2, 3\}$ is the set $\{1, 3\}$; that is, $\{1, 3, 5\} \cap \{1, 2, 3\} = \{1, 3\}$. ◀

EXAMPLE 4 The intersection of the set of all computer science majors at your school and the set of all mathematics majors is the set of all students who are joint majors in mathematics and computer science. ◀

DEFINITION 3 Two sets are called *disjoint* if their intersection is the empty set.

EXAMPLE 5 Let $A = \{1, 3, 5, 7, 9\}$ and $B = \{2, 4, 6, 8, 10\}$. Because $A \cap B = \emptyset$, A and B are disjoint. ◀

We often are interested in finding the cardinality of a union of two finite sets A and B. Note that $|A| + |B|$ counts each element that is in A but not in B or in B but not in A exactly once, and each element that is in both A and B exactly twice. Thus, if the number of elements that are in both A and B is subtracted from $|A| + |B|$, elements in $A \cap B$ will be counted only once. Hence,

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

The generalization of this result to unions of an arbitrary number of sets is called the **principle of inclusion–exclusion.** The principle of inclusion–exclusion is an important technique used in enumeration. We will discuss this principle and other counting techniques in detail in Chapters 5 and 7.

There are other important ways to combine sets.

$A - B$ is shaded.

FIGURE 3 Venn Diagram for the Difference of A and B.

$\overline{A}$ is shaded.

FIGURE 4 Venn Diagram for the Complement of the Set A.

DEFINITION 4 Let A and B be sets. The *difference* of A and B, denoted by $A - B$, is the set containing those elements that are in A but not in B. The difference of A and B is also called the *complement of B with respect to A.*

An element x belongs to the difference of A and B if and only if $x \in A$ and $x \notin B$. This tells us that

$$A - B = \{x \mid x \in A \wedge x \notin B\}.$$

The Venn diagram shown in Figure 3 represents the difference of the sets A and B. The shaded area inside the circle that represents A and outside the circle that represents B is the area that represents $A - B$.

We give some examples of differences of sets.

EXAMPLE 6 The difference of $\{1, 3, 5\}$ and $\{1, 2, 3\}$ is the set $\{5\}$; that is, $\{1, 3, 5\} - \{1, 2, 3\} = \{5\}$. This is different from the difference of $\{1, 2, 3\}$ and $\{1, 3, 5\}$, which is the set $\{2\}$. ◀

EXAMPLE 7 The difference of the set of computer science majors at your school and the set of mathematics majors at your school is the set of all computer science majors at your school who are not also mathematics majors. ◀

Once the universal set U has been specified, the **complement** of a set can be defined.

DEFINITION 5 Let U be the universal set. The *complement* of the set A, denoted by $\overline{A}$, is the complement of A with respect to U. In other words, the complement of the set A is $U - A$.

An element belongs to $\overline{A}$ if and only if $x \notin A$. This tells us that

$$\overline{A} = \{x \mid x \notin A\}.$$

In Figure 4 the shaded area outside the circle representing A is the area representing $\overline{A}$.

We give some examples of the complement of a set.

TABLE 1 Set Identities.

Identity	*Name*
$A \cup \emptyset = A$ $A \cap U = A$	Identity laws
$A \cup U = U$ $A \cap \emptyset = \emptyset$	Domination laws
$A \cup A = A$ $A \cap A = A$	Idempotent laws
$\overline{(\overline{A})} = A$	Complementation law
$A \cup B = B \cup A$ $A \cap B = B \cap A$	Commutative laws
$A \cup (B \cup C) = (A \cup B) \cup C$ $A \cap (B \cap C) = (A \cap B) \cap C$	Associative laws
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	Distributive laws
$\overline{A \cup B} = \overline{A} \cap \overline{B}$ $\overline{A \cap B} = \overline{A} \cup \overline{B}$	De Morgan's laws
$A \cup (A \cap B) = A$ $A \cap (A \cup B) = A$	Absorption laws
$A \cup \overline{A} = U$ $A \cap \overline{A} = \emptyset$	Complement laws

EXAMPLE 8 Let $A = \{a, e, i, o, u\}$ (where the universal set is the set of letters of the English alphabet). Then $\overline{A} = \{b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, y, z\}$. ◀

EXAMPLE 9 Let A be the set of positive integers greater than 10 (with universal set the set of all positive integers). Then $\overline{A} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. ◀

Set Identities

Table 1 lists the most important set identities. We will prove several of these identities here, using three different methods. These methods are presented to illustrate that there are often many different approaches to the solution of a problem. The proofs of the remaining identities will be left as exercises. The reader should note the similarity between these set identities and the logical equivalences discussed in Section 1.2. In fact, the set identities given can be proved directly from the corresponding logical equivalences. Furthermore, both are special cases of identities that hold for Boolean algebra (discussed in Chapter 11).

One way to show that two sets are equal is to show that each is a subset of the other. Recall that to show that one set is a subset of a second set, we can show that if an element belongs to the first set, then it must also belong to the second set. We generally use a direct proof to do this. We illustrate this type of proof by establishing the second of De Morgan's laws.

EXAMPLE 10 Prove that $\overline{A \cap B} = \overline{A} \cup \overline{B}$.

Solution: To show that $\overline{A \cap B} = \overline{A} \cup \overline{B}$, we will show that $\overline{A \cap B} \subseteq \overline{A} \cup \overline{B}$ and that $\overline{A} \cup \overline{B} \subseteq \overline{A \cap B}$.

Extra Examples

First, we will show that $\overline{A \cap B} \subseteq \overline{A} \cup \overline{B}$. So suppose that $x \in \overline{A \cap B}$. By the definition of complement, $x \notin A \cap B$. By the definition of intersection, $\neg((x \in A) \wedge (x \in B))$ is true. Applying De Morgan's law (from logic), we see that $\neg(x \in A)$ or $\neg(x \in B)$. Hence, by the definition of negation, $x \notin A$ or $x \notin B$. By the definition of complement, $x \in \overline{A}$ or $x \in \overline{B}$. It follows by the definition of union that $x \in \overline{A} \cup \overline{B}$. This shows that $\overline{A \cap B} \subseteq \overline{A} \cup \overline{B}$.

Next, we will show that $\overline{A} \cup \overline{B} \subseteq \overline{A \cap B}$. Now suppose that $x \in \overline{A} \cup \overline{B}$. By the definition of union, $x \in \overline{A}$ or $x \in \overline{B}$. Using the definition of complement, we see that $x \notin A$ or $x \notin B$. Consequently, $\neg(x \in A) \vee \neg(x \in B)$ is true. By De Morgan's law (from logic), we conclude that $\neg((x \in A) \wedge (x \in B))$ is true. By the definition of intersection, it follows that $\neg(x \in A \cap B)$ holds. We use the definition of complement to conclude that $x \in \overline{A \cap B}$. This shows that $\overline{A} \cup \overline{B} \subseteq \overline{A \cap B}$. Because we have shown that each set is a subset of the other, the two sets are equal, and the identity is proved. ◀

We can more succinctly express the reasoning used in Example 10 using set builder notation, as Example 11 illustrates.

EXAMPLE 11 Use set builder notation and logical equivalences to establish the second De Morgan law $\overline{A \cap B} = \overline{A} \cup \overline{B}$.

Solution: We can prove this identity with the following steps.

$\overline{A \cap B} = \{x \mid x \notin A \cap B\}$	by definition of complement
$= \{x \mid \neg(x \in (A \cap B))\}$	by definition of does not belong symbol
$= \{x \mid \neg(x \in A \wedge x \in B)\}$	by definition of intersection
$= \{x \mid \neg(x \in A) \vee \neg(x \in B)\}$	by the first De Morgan law for logical equivalences
$= \{x \mid x \notin A \vee x \notin B\}$	by definition of does not belong symbol
$= \{x \mid x \in \overline{A} \vee x \in \overline{B}\}$	by definition of complement
$= \{x \mid x \in \overline{A} \cup \overline{B}\}$	by definition of union
$= \overline{A} \cup \overline{B}$	by meaning of set builder notation

Note that besides the definitions of complement, union, set membership, and set builder notation, this proof uses the first De Morgan law for logical equivalences. ◀

Proving a set identity involving more than two sets by showing each side of the identity is a subset of the other often requires that we keep track of different cases, as illustrated by the proof in Example 12 of one of the distributive laws for sets.

EXAMPLE 12 Prove the first distributive law from Table 1, which states that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ for all sets A, B, and C.

Solution: We will prove this identity by showing that each side is a subset of the other side.

Suppose that $x \in A \cap (B \cup C)$. Then $x \in A$ and $x \in B \cup C$. By the definition of union, it follows that $x \in A$, and $x \in B$ or $x \in C$ (or both). Consequently, we know that $x \in A$ and $x \in B$ or that $x \in A$ and $x \in C$. By the definition of intersection, it follows that $x \in A \cap B$ or $x \in A \cap C$. Using the definition of union, we conclude that $x \in (A \cap B) \cup (A \cap C)$. We conclude that $A \cap (B \cup C) \subseteq (A \cap B) \cup (A \cap C)$.

TABLE 2 A Membership Table for the Distributive Property.

A	B	C	$B \cup C$	$A \cap (B \cup C)$	$A \cap B$	$A \cap C$	$(A \cap B) \cup (A \cap C)$
1	1	1	1	1	1	1	1
1	1	0	1	1	1	0	1
1	0	1	1	1	0	1	1
1	0	0	0	0	0	0	0
0	1	1	1	0	0	0	0
0	1	0	1	0	0	0	0
0	0	1	1	0	0	0	0
0	0	0	0	0	0	0	0

Now suppose that $x \in (A \cap B) \cup (A \cap C)$. Then, by the definition of union, $x \in A \cap B$ or $x \in A \cap C$. By the definition of intersection, it follows that $x \in A$ and $x \in B$ or that $x \in A$ and $x \in C$. From this we see that $x \in A$, and $x \in B$ or $x \in C$. Consequently, by the definition of union we see that $x \in A$ and $x \in B \cup C$. Furthermore, by the definition of intersection, it follows that $x \in A \cap (B \cup C)$. We conclude that $(A \cap B) \cup (A \cap C) \subseteq A \cap (B \cup C)$. This completes the proof of the identity. ◀

Set identities can also be proved using **membership tables.** We consider each combination of sets that an element can belong to and verify that elements in the same combinations of sets belong to both the sets in the identity. To indicate that an element is in a set, a 1 is used; to indicate that an element is not in a set, a 0 is used. (The reader should note the similarity between membership tables and truth tables.)

EXAMPLE 13 Use a membership table to show that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Solution: The membership table for these combinations of sets is shown in Table 2. This table has eight rows. Because the columns for $A \cap (B \cup C)$ and $(A \cap B) \cup (A \cap C)$ are the same, the identity is valid. ◀

Additional set identities can be established using those that we have already proved. Consider Example 14.

EXAMPLE 14 Let A, B, and C be sets. Show that

$$\overline{A \cup (B \cap C)} = (\overline{C} \cup \overline{B}) \cap \overline{A}.$$

Solution: We have

$$\begin{aligned}
\overline{A \cup (B \cap C)} &= \overline{A} \cap (\overline{B \cap C}) && \text{by the first De Morgan law} \\
&= \overline{A} \cap (\overline{B} \cup \overline{C}) && \text{by the second De Morgan law} \\
&= (\overline{B} \cup \overline{C}) \cap \overline{A} && \text{by the commutative law for intersections} \\
&= (\overline{C} \cup \overline{B}) \cap \overline{A} && \text{by the commutative law for unions.}
\end{aligned}$$

◀

Generalized Unions and Intersections

Because unions and intersections of sets satisfy associative laws, the sets $A \cup B \cup C$ and $A \cap B \cap C$ are well defined; that is, the meaning of this notation is unambiguous when A, B, and

(a) $A \cup B \cup C$ is shaded. (b) $A \cap B \cap C$ is shaded.

FIGURE 5 The Union and Intersection of A, B, and C.

C are sets. That is, we do not have to use parentheses to indicate which operation comes first because $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$. Note that $A \cup B \cup C$ contains those elements that are in at least one of the sets A, B, and C, and that $A \cap B \cap C$ contains those elements that are in all of A, B, and C. These combinations of the three sets, A, B, and C, are shown in Figure 5.

EXAMPLE 15 Let $A = \{0, 2, 4, 6, 8\}$, $B = \{0, 1, 2, 3, 4\}$, and $C = \{0, 3, 6, 9\}$. What are $A \cup B \cup C$ and $A \cap B \cap C$?

Solution: The set $A \cup B \cup C$ contains those elements in at least one of A, B, and C. Hence,

$$A \cup B \cup C = \{0, 1, 2, 3, 4, 6, 8, 9\}.$$

The set $A \cap B \cap C$ contains those elements in all three of A, B, and C. Thus,

$$A \cap B \cap C = \{0\}.$$ ◀

We can also consider unions and intersections of an arbitrary number of sets. We use these definitions.

DEFINITION 6 The *union* of a collection of sets is the set that contains those elements that are members of at least one set in the collection.

We use the notation

$$A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{i=1}^{n} A_i$$

to denote the union of the sets $A_1, A_2, \ldots, A_n$.

DEFINITION 7 The *intersection* of a collection of sets is the set that contains those elements that are members of all the sets in the collection.

We use the notation

$$A_1 \cap A_2 \cap \cdots \cap A_n = \bigcap_{i=1}^{n} A_i$$

to denote the intersection of the sets $A_1, A_2, \ldots, A_n$. We illustrate generalized unions and intersections with Example 16.

EXAMPLE 16 Let $A_i = \{i, i+1, i+2, \ldots\}$. Then,

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} \{i, i+1, i+2, \ldots\} = \{1, 2, 3, \ldots\},$$

and

$$\bigcap_{i=1}^{n} A_i = \bigcap_{i=1}^{n} \{i, i+1, i+2, \ldots\} = \{n, n+1, n+2, \ldots\}.$$ ◀

We can extend the notation we have introduce for unions and intersections to other families of sets. In particular, we can use the notation

$$A_1 \cup A_2 \cup \cdots \cup A_n \cup \cdots = \bigcup_{i=1}^{\infty} A_i$$

to denote the union of the sets $A_1, A_2, \ldots, A_n, \ldots$. Similarly, the intersection of these sets can be denoted by

$$A_1 \cap A_2 \cap \cdots \cap A_n \cap \cdots = \bigcap_{i=1}^{\infty} A_i.$$

More generally, the notations $\bigcap_{i \in I} A_i$ and $\bigcup_{i \in I} A_i$ are used to denote the intersection and union of the sets A_i for $i \in I$, respectively. Note that we have $\bigcap_{i \in I} A_i = \{x \mid \forall i \in I\ (x \in A_i)\}$ and $\bigcup_{i \in I} A_i = \{x \mid \exists i \in I\ (x \in A_i)\}$.

EXAMPLE 17 Suppose that $A_i = \{1, 2, 3, \ldots, i\}$ for $i = 1, 2, 3, \ldots$. Then,

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} \{1, 2, 3, \ldots, i\} = \{1, 2, 3, \ldots\} = \mathbf{Z}^+$$

and

$$\bigcap_{i=1}^{\infty} A_i = \bigcap_{i=1}^{\infty} \{1, 2, 3, \ldots, i\} = \{1\}.$$

To see that the union of these sets is the set of positive integers, note that every positive integer is in at least one of the sets, because the integer n belongs to $A_n = \{1, 2, \ldots, n\}$ and every element of the sets in the union is a positive integer. To see that the intersection of these sets, note that the only element that belongs to all the sets $A_1, A_2, \ldots$ is 1. To see this note that $A_1 = \{1\}$ and $1 \in A_i$ for $i = 1, 2, \ldots$. ◀

Computer Representation of Sets

There are various ways to represent sets using a computer. One method is to store the elements of the set in an unordered fashion. However, if this is done, the operations of computing the union, intersection, or difference of two sets would be time-consuming, because each of these operations would require a large amount of searching for elements. We will present a method for storing elements using an arbitrary ordering of the elements of the universal set. This method of representing sets makes computing combinations of sets easy.

Assume that the universal set U is finite (and of reasonable size so that the number of elements of U is not larger than the memory size of the computer being used). First, specify an arbitrary ordering of the elements of U, for instance $a_1, a_2, \ldots, a_n$. Represent a subset A of U with the bit string of length n, where the ith bit in this string is 1 if a_i belongs to A and is 0 if a_i does not belong to A. Example 18 illustrates this technique.

EXAMPLE 18 Let $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, and the ordering of elements of U has the elements in increasing order; that is, $a_i = i$. What bit strings represent the subset of all odd integers in U, the subset of all even integers in U, and the subset of integers not exceeding 5 in U?

Solution: The bit string that represents the set of odd integers in U, namely, $\{1, 3, 5, 7, 9\}$, has a one bit in the first, third, fifth, seventh, and ninth positions, and a zero elsewhere. It is

10 1010 1010.

(We have split this bit string of length ten into blocks of length four for easy reading because long bit strings are difficult to read.) Similarly, we represent the subset of all even integers in U, namely, $\{2, 4, 6, 8, 10\}$, by the string

01 0101 0101.

The set of all integers in U that do not exceed 5, namely, $\{1, 2, 3, 4, 5\}$, is represented by the string

11 1110 0000. ◀

Using bit strings to represent sets, it is easy to find complements of sets and unions, intersections, and differences of sets. To find the bit string for the complement of a set from the bit string for that set, we simply change each 1 to a 0 and each 0 to 1, because $x \in A$ if and only if $x \notin \overline{A}$. Note that this operation corresponds to taking the negation of each bit when we associate a bit with a truth value—with 1 representing true and 0 representing false.

EXAMPLE 19 We have seen that the bit string for the set $\{1, 3, 5, 7, 9\}$ (with universal set $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$) is

10 1010 1010.

What is the bit string for the complement of this set?

Solution: The bit string for the complement of this set is obtained by replacing 0s with 1s and vice versa. This yields the string

01 0101 0101,

which corresponds to the set $\{2, 4, 6, 8, 10\}$. ◀

To obtain the bit string for the union and intersection of two sets we perform bitwise Boolean operations on the bit strings representing the two sets. The bit in the ith position of the bit string of the union is 1 if either of the bits in the ith position in the two strings is 1 (or both are 1), and is 0 when both bits are 0. Hence, the bit string for the union is the bitwise *OR* of the bit strings for the two sets. The bit in the ith position of the bit string of the intersection is 1 when the bits in the corresponding position in the two strings are both 1, and is 0 when either of the two bits is 0 (or both are). Hence, the bit string for the intersection is the bitwise *AND* of the bit strings for the two sets.

EXAMPLE 20 The bit strings for the sets $\{1, 2, 3, 4, 5\}$ and $\{1, 3, 5, 7, 9\}$ are 11 1110 0000 and 10 1010 1010, respectively. Use bit strings to find the union and intersection of these sets.

Solution: The bit string for the union of these sets is

$$11\ 1110\ 0000 \vee 10\ 1010\ 1010 = 11\ 1110\ 1010,$$

which corresponds to the set $\{1, 2, 3, 4, 5, 7, 9\}$. The bit string for the intersection of these sets is

$$11\ 1110\ 0000 \wedge 10\ 1010\ 1010 = 10\ 1010\ 0000,$$

which corresponds to the set $\{1, 3, 5\}$. ◀

Exercises

1. Let A be the set of students who live within one mile of school and let B be the set of students who walk to classes. Describe the students in each of these sets.
   a) $A \cap B$ b) $A \cup B$
   c) $A - B$ d) $B - A$
2. Suppose that A is the set of sophomores at your school and B is the set of students in discrete mathematics at your school. Express each of these sets in terms of A and B.
   a) the set of sophomores taking discrete mathematics in your school
   b) the set of sophomores at your school who are not taking discrete mathematics
   c) the set of students at your school who either are sophomores or are taking discrete mathematics
   d) the set of students at your school who either are not sophomores or are not taking discrete mathematics
3. Let $A = \{1, 2, 3, 4, 5\}$ and $B = \{0, 3, 6\}$. Find
   a) $A \cup B$. b) $A \cap B$.
   c) $A - B$. d) $B - A$.
4. Let $A = \{a, b, c, d, e\}$ and $B = \{a, b, c, d, e, f, g, h\}$. Find
   a) $A \cup B$. b) $A \cap B$.
   c) $A - B$. d) $B - A$.

In Exercises 5–10 assume that A is a subset of some underlying universal set U.

5. Prove the complementation law in Table 1 by showing that $\overline{\overline{A}} = A$.
6. Prove the identity laws in Table 1 by showing that
   a) $A \cup \emptyset = A$. b) $A \cap U = A$.
7. Prove the domination laws in Table 1 by showing that
   a) $A \cup U = U$. b) $A \cap \emptyset = \emptyset$.
8. Prove the idempotent laws in Table 1 by showing that
   a) $A \cup A = A$. b) $A \cap A = A$.
9. Prove the complement laws in Table 1 by showing that
   a) $A \cup \overline{A} = U$. b) $A \cap \overline{A} = \emptyset$.
10. Show that
    a) $A - \emptyset = A$. b) $\emptyset - A = \emptyset$.
11. Let A and B be sets. Prove the commutative laws from Table 1 by showing that
    a) $A \cup B = B \cup A$.
    b) $A \cap B = B \cap A$.
12. Prove the first absorption law from Table 1 by showing that if A and B are sets, then $A \cup (A \cap B) = A$.
13. Prove the second absorption law from Table 1 by showing that if A and B are sets, then $A \cap (A \cup B) = A$.
14. Find the sets A and B if $A - B = \{1, 5, 7, 8\}$, $B - A = \{2, 10\}$, and $A \cap B = \{3, 6, 9\}$.
15. Prove the first De Morgan law in Table 1 by showing that if A and B are sets, then $\overline{A \cup B} = \overline{A} \cap \overline{B}$

a) by showing each side is a subset of the other side.
b) using a membership table.

16. Let A and B be sets. Show that
a) $(A \cap B) \subseteq A$. b) $A \subseteq (A \cup B)$.
c) $A - B \subseteq A$. d) $A \cap (B - A) = \emptyset$.
e) $A \cup (B - A) = A \cup B$.

17. Show that if A, B, and C are sets, then $\overline{A \cap B \cap C} = \overline{A} \cup \overline{B} \cup \overline{C}$
a) by showing each side is a subset of the other side.
b) using a membership table.

18. Let A, B, and C be sets. Show that
a) $(A \cup B) \subseteq (A \cup B \cup C)$.
b) $(A \cap B \cap C) \subseteq (A \cap B)$.
c) $(A - B) - C \subseteq A - C$.
d) $(A - C) \cap (C - B) = \emptyset$.
e) $(B - A) \cup (C - A) = (B \cup C) - A$.

19. Show that if A and B are sets, then $A - B = A \cap \overline{B}$.

20. Show that if A and B are sets, then $(A \cap B) \cup (A \cap \overline{B}) = A$.

21. Prove the first associative law from Table 1 by showing that if A, B, and C are sets, then $A \cup (B \cup C) = (A \cup B) \cup C$.

22. Prove the second associative law from Table 1 by showing that if A, B, and C are sets, then $A \cap (B \cap C) = (A \cap B) \cap C$.

23. Prove the second distributive law from Table 1 by showing that if A, B, and C are sets, then $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

24. Let A, B, and C be sets. Show that $(A - B) - C = (A - C) - (B - C)$.

25. Let $A = \{0, 2, 4, 6, 8, 10\}$, $B = \{0, 1, 2, 3, 4, 5, 6\}$, and $C = \{4, 5, 6, 7, 8, 9, 10\}$. Find
a) $A \cap B \cap C$. b) $A \cup B \cup C$.
c) $(A \cup B) \cap C$. d) $(A \cap B) \cup C$.

26. Draw the Venn diagrams for each of these combinations of the sets A, B, and C.
a) $A \cap (B \cup C)$ b) $\overline{A} \cap \overline{B} \cap \overline{C}$
c) $(A - B) \cup (A - C) \cup (B - C)$

27. Draw the Venn diagrams for each of these combinations of the sets A, B, and C.
a) $A \cap (B - C)$ b) $(A \cap B) \cup (A \cap C)$
c) $(A \cap \overline{B}) \cup (A \cap \overline{C})$

28. Draw the Venn diagrams for each of these combinations of the sets A, B, C, and D.
a) $(A \cap B) \cup (C \cap D)$ b) $\overline{A} \cup \overline{B} \cup \overline{C} \cup \overline{D}$
c) $A - (B \cap C \cap D)$

29. What can you say about the sets A and B if we know that
a) $A \cup B = A$? b) $A \cap B = A$?
c) $A - B = A$? d) $A \cap B = B \cap A$?
e) $A - B = B - A$?

30. Can you conclude that $A = B$ if A, B, and C are sets such that
a) $A \cup C = B \cup C$? b) $A \cap C = B \cap C$?
c) $A \cup B = B \cup C$ and $A \cap C = B \cap C$?

31. Let A and B be subsets of a universal set U. Show that $A \subseteq B$ if and only if $\overline{B} \subseteq \overline{A}$?

The **symmetric difference** of A and B, denoted by $A \oplus B$, is the set containing those elements in either A or B, but not in both A and B.

32. Find the symmetric difference of $\{1, 3, 5\}$ and $\{1, 2, 3\}$.

33. Find the symmetric difference of the set of computer science majors at a school and the set of mathematics majors at this school.

34. Draw a Venn diagram for the symmetric difference of the sets A and B.

35. Show that $A \oplus B = (A \cup B) - (A \cap B)$.

36. Show that $A \oplus B = (A - B) \cup (B - A)$.

37. Show that if A is a subset of a universal set U, then
a) $A \oplus A = \emptyset$. b) $A \oplus \emptyset = A$.
c) $A \oplus U = \overline{A}$. d) $A \oplus \overline{A} = U$.

38. Show that if A and B are sets, then
a) $A \oplus B = B \oplus A$. b) $(A \oplus B) \oplus B = A$.

39. What can you say about the sets A and B if $A \oplus B = A$?

***40.** Determine whether the symmetric difference is associative; that is, if A, B, and C are sets, does it follow that $A \oplus (B \oplus C) = (A \oplus B) \oplus C$?

***41.** Suppose that A, B, and C are sets such that $A \oplus C = B \oplus C$. Must it be the case that $A = B$?

42. If A, B, C, and D are sets, does it follow that $(A \oplus B) \oplus (C \oplus D) = (A \oplus C) \oplus (B \oplus D)$?

43. If A, B, C, and D are sets, does it follow that $(A \oplus B) \oplus (C \oplus D) = (A \oplus D) \oplus (B \oplus C)$?

***44.** Show that if A, B, and C are sets, then

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

(This is a special case of the inclusion–exclusion principle, which will be studied in Chapter 7.)

***45.** Let $A_i = \{1, 2, 3, \ldots, i\}$ for $i = 1, 2, 3, \ldots$. Find
a) $\bigcup_{i=1}^{n} A_i$. b) $\bigcap_{i=1}^{n} A_i$.

***46.** Let $A_i = \{\ldots, -2, -1, 0, 1, \ldots, i\}$. Find
a) $\bigcup_{i=1}^{n} A_i$. b) $\bigcap_{i=1}^{n} A_i$.

47. Let A_i be the set of all nonempty bit strings (that is, bit strings of length at least one) of length not exceeding i. Find
a) $\bigcup_{i=1}^{n} A_i$. b) $\bigcap_{i=1}^{n} A_i$.

48. Find $\bigcup_{i=1}^{\infty} A_i$ and $\bigcap_{i=1}^{\infty} A_i$ if for every positive integer i,
a) $A_i = \{i, i+1, i+2, \ldots\}$.
b) $A_i = \{0, i\}$.
c) $A_i = (0, i)$, that is, the set of real numbers x with $0 < x < i$.
d) $A_i = (i, \infty)$, that is, the set of real numbers x with $x > i$.

49. Find $\bigcup_{i=1}^{\infty} A_i$ and $\bigcap_{i=1}^{\infty} A_i$ if for every positive integer i,

a) $A_i = \{-i, -i+1, \ldots, -1, 0, 1, \ldots, i-1, i\}$.
b) $A_i = \{-i, i\}$.
c) $A_i = [-i, i]$, that is, the set of real numbers x with $-i \le x \le i$.
d) $A_i = [i, \infty]$, that is, the set of real numbers x with $x \ge i$.

50. Suppose that the universal set is $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Express each of these sets with bit strings where the ith bit in the string is 1 if i is in the set and 0 otherwise.

a) $\{3, 4, 5\}$
b) $\{1, 3, 6, 10\}$
c) $\{2, 3, 4, 7, 8, 9\}$

51. Using the same universal set as in the last problem, find the set specified by each of these bit strings.

a) 11 1100 1111
b) 01 0111 1000
c) 10 0000 0001

52. What subsets of a finite universal set do these bit strings represent?

a) the string with all zeros
b) the string with all ones

53. What is the bit string corresponding to the difference of two sets?

54. What is the bit string corresponding to the symmetric difference of two sets?

55. Show how bitwise operations on bit strings can be used to find these combinations of $A = \{a, b, c, d, e\}$, $B = \{b, c, d, g, p, t, v\}$, $C = \{c, e, i, o, u, x, y, z\}$, and $D = \{d, e, h, i, n, o, t, u, x, y\}$.

a) $A \cup B$ **b)** $A \cap B$
c) $(A \cup D) \cap (B \cup C)$ **d)** $A \cup B \cup C \cup D$

56. How can the union and intersection of n sets that all are subsets of the universal set U be found using bit strings?

The **successor** of the set A is the set $A \cup \{A\}$.

57. Find the successors of the following sets.

a) $\{1, 2, 3\}$ **b)** $\emptyset$
c) $\{\emptyset\}$ **d)** $\{\emptyset, \{\emptyset\}\}$

58. How many elements does the successor of a set with n elements have?

Sometimes the number of times that an element occurs in an unordered collection matters. **Multisets** are unordered collections of elements where an element can occur as a member more than once. The notation $\{m_1 \cdot a_1, m_2 \cdot a_2, \ldots, m_r \cdot a_r\}$ denotes the multiset with element a_1 occurring m_1 times, element a_2 occurring m_2 times, and so on. The numbers m_i, $i = 1, 2, \ldots, r$ are called the **multiplicities** of the elements a_i, $i = 1, 2, \ldots, r$.

Let P and Q be multisets. The **union** of the multisets P and Q is the multiset where the multiplicity of an element is the maximum of its multiplicities in P and Q. The **intersection** of P and Q is the multiset where the multiplicity of an element is the minimum of its multiplicities in P and Q. The **difference** of P and Q is the multiset where the multiplicity of an element is the multiplicity of the element in P less its multiplicity in Q unless this difference is negative, in which case the multiplicity is 0. The **sum** of P and Q is the multiset where the multiplicity of an element is the sum of multiplicities in P and Q. The union, intersection, and difference of P and Q are denoted by $P \cup Q$, $P \cap Q$, and $P - Q$, respectively (where these operations should not be confused with the analogous operations for sets). The sum of P and Q is denoted by $P + Q$.

59. Let A and B be the multisets $\{3 \cdot a, 2 \cdot b, 1 \cdot c\}$ and $\{2 \cdot a, 3 \cdot b, 4 \cdot d\}$, respectively. Find

a) $A \cup B$. **b)** $A \cap B$.
c) $A - B$. **d)** $B - A$.
e) $A + B$.

60. Suppose that A is the multiset that has as its elements the types of computer equipment needed by one department of a university where the multiplicities are the number of pieces of each type needed, and B is the analogous multiset for a second department of the university. For instance, A could be the multiset {107 · personal computers, 44 · routers, 6 · servers} and B could be the multiset {14 · personal computers, 6 · routers, 2 · mainframes}.

a) What combination of A and B represents the equipment the university should buy assuming both departments use the same equipment?
b) What combination of A and B represents the equipment that will be used by both departments if both departments use the same equipment?
c) What combination of A and B represents the equipment that the second department uses, but the first department does not, if both departments use the same equipment?
d) What combination of A and B represents the equipment that the university should purchase if the departments do not share equipment?

Fuzzy sets are used in artificial intelligence. Each element in the universal set U has a **degree of membership,** which is a real number between 0 and 1 (including 0 and 1), in a fuzzy set S. The fuzzy set S is denoted by listing the elements with their degrees of membership (elements with 0 degree of membership are not listed). For instance, we write {0.6 Alice, 0.9 Brian, 0.4 Fred, 0.1 Oscar, 0.5 Rita} for the set F (of famous people) to indicate that Alice has a 0.6 degree of membership in F, Brian has a 0.9 degree of membership in F, Fred has a 0.4 degree of membership in F, Oscar has a 0.1 degree of membership in F, and Rita has a 0.5 degree of membership in F (so that Brian is the most famous and Oscar is the least famous of these people). Also suppose that R is the set of rich people with R = {0.4 Alice, 0.8 Brian, 0.2 Fred, 0.9 Oscar, 0.7 Rita}.

61. The **complement** of a fuzzy set S is the set $\overline{S}$, with the degree of the membership of an element in $\overline{S}$ equal to 1 minus the degree of membership of this element in S. Find $\overline{F}$ (the fuzzy set of people who are not famous) and $\overline{R}$ (the fuzzy set of people who are not rich).

62. The **union** of two fuzzy sets S and T is the fuzzy set $S \cup T$, where the degree of membership of an element in $S \cup T$ is the maximum of the degrees of membership of this element in S and in T. Find the fuzzy set $F \cup R$ of rich or famous people.

63. The **intersection** of two fuzzy sets S and T is the fuzzy set $S \cap T$, where the degree of membership of an element in $S \cap T$ is the minimum of the degrees of membership of this element in S and in T. Find the fuzzy set $F \cap R$ of rich and famous people.

2.3 Functions

Introduction

In many instances we assign to each element of a set a particular element of a second set (which may be the same as the first). For example, suppose that each student in a discrete mathematics class is assigned a letter grade from the set $\{A, B, C, D, F\}$. And suppose that the grades are A for Adams, C for Chou, B for Goodfriend, A for Rodriguez, and F for Stevens. This assignment of grades is illustrated in Figure 1.

This assignment is an example of a function. The concept of a function is extremely important in mathematics and computer science. For example, in discrete mathematics functions are used in the definition of such discrete structures as sequences and strings. Functions are also used to represent how long it takes a computer to solve problems of a given size. Many computer programs and subroutines are designed to calculate values of functions. Recursive functions, which are functions defined in terms of themselves, are used throughout computer science; they will be studied in Chapter 4. This section reviews the basic concepts involving functions needed in discrete mathematics.

DEFINITION 1 Let A and B be nonempty sets. A *function* f from A to B is an assignment of exactly one element of B to each element of A. We write $f(a) = b$ if b is the unique element of B assigned by the function f to the element a of A. If f is a function from A to B, we write $f : A \to B$.

Remark: Functions are sometimes also called **mappings** or **transformations.**

Assessment

Functions are specified in many different ways. Sometimes we explicitly state the assignments, as in Figure 1. Often we give a formula, such as $f(x) = x + 1$, to define a function. Other times we use a computer program to specify a function.

FIGURE 1 Assignment of Grades in a Discrete Mathematics Class.

FIGURE 2 The Function f Maps A to B.

A function $f : A \to B$ can also be defined in terms of a relation from A to B. Recall from Section 2.1 that a relation from A to B is just a subset of $A \times B$. A relation from A to B that contains one, and only one, ordered pair (a, b) for every element $a \in A$, defines a function f from A to B. This function is defined by the assignment $f(a) = b$, where (a, b) is the unique ordered pair in the relation that has a as its first element.

DEFINITION 2 If f is a function from A to B, we say that A is the *domain* of f and B is the *codomain* of f. If $f(a) = b$, we say that b is the *image* of a and a is a *preimage* of b. The *range* of f is the set of all images of elements of A. Also, if f is a function from A to B, we say that *f maps A to B*.

Figure 2 represents a function f from A to B.

When we define a function we specify its domain, its codomain, and the mapping of elements of the domain to elements in the codomain. Two functions are **equal** when they have the same domain, have the same codomain, and map elements of their common domain to the same elements in their common codomain. Note that if we change either the domain or the codomain of a function, then we obtain a different function. If we change the mapping of elements, then we also obtain a different function.

Examples 1–5 provide examples of functions. In each case, we describe the domain, the codomain, the range, and the assignment of values to elements of the domain.

EXAMPLE 1 What are the domain, codomain, and range of the function that assigns grades to students described in the first paragraph of the introduction of this section?

Solution: Let G be the function that assigns a grade to a student in our discrete mathematics class. Note that $G(\text{Adams}) = A$, for instance. The domain of G is the set {Adams, Chou, Goodfriend, Rodriguez, Stevens}, and the codomain is the set $\{A, B, C, D, F\}$. The range of G is the set $\{A, B, C, F\}$, because each grade except D is assigned to some student. ◀

EXAMPLE 2 Let R be the relation consisting of ordered pairs (Abdul, 22), (Brenda, 24), (Carla, 21), (Desire, 22), (Eddie, 24), and (Felicia, 22), where each pair consists of a graduate student and the age of this student. What is the function that this relation determines?

Solution: This relation defines the function f, where with $f(\text{Abdul}) = 22$, $f(\text{Brenda}) = 24$, $f(\text{Carla}) = 21$, $f(\text{Desire}) = 22$, $f(\text{Eddie}) = 24$, and $f(\text{Felicia}) = 22$. Here the domain is the set {Abdul, Brenda, Carla, Desire, Eddie, Felicia}. To define the function f, we need to specify a codomain. Here, we can take the codomain to be the set of positive integers to make sure

that the codomain contains all possible ages of students. (Note that we could choose a smaller codomain, but that would change the function.) Finally, the range is the set $\{21, 22, 24\}$. ◀

EXAMPLE 3 Let f be the function that assigns the last two bits of a bit string of length 2 or greater to that string. For example, $f(11010) = 10$. Then, the domain of f is the set of all bit strings of length 2 or greater, and both the codomain and range are the set $\{00, 01, 10, 11\}$. ◀

Extra Examples

EXAMPLE 4 Let $f\colon \mathbf{Z} \to \mathbf{Z}$ assign the square of an integer to this integer. Then, $f(x) = x^2$, where the domain of f is the set of all integers, we take the the codomain of f to be the set of all integers, and the range of f is the set of all integers that are perfect squares, namely, $\{0, 1, 4, 9, \ldots\}$. ◀

EXAMPLE 5 The domain and codomain of functions are often specified in programming languages. For instance, the Java statement

int **floor**(float real){...}

and the Pascal statement

function *floor*(x: **real**): **integer**

both state that the domain of the floor function is the set of real numbers and its codomain is the set of integers. ◀

Two real-valued functions with the same domain can be added and multiplied.

DEFINITION 3 Let f_1 and f_2 be functions from A to $\mathbf{R}$. Then $f_1 + f_2$ and $f_1 f_2$ are also functions from A to $\mathbf{R}$ defined by

$$(f_1 + f_2)(x) = f_1(x) + f_2(x),$$
$$(f_1 f_2)(x) = f_1(x) f_2(x).$$

Note that the functions $f_1 + f_2$ and $f_1 f_2$ have been defined by specifying their values at x in terms of the values of f_1 and f_2 at x.

EXAMPLE 6 Let f_1 and f_2 be functions from $\mathbf{R}$ to $\mathbf{R}$ such that $f_1(x) = x^2$ and $f_2(x) = x - x^2$. What are the functions $f_1 + f_2$ and $f_1 f_2$?

Solution: From the definition of the sum and product of functions, it follows that

$$(f_1 + f_2)(x) = f_1(x) + f_2(x) = x^2 + (x - x^2) = x$$

and

$$(f_1 f_2)(x) = x^2(x - x^2) = x^3 - x^4.$$ ◀

When f is a function from a set A to a set B, the image of a subset of A can also be defined.

DEFINITION 4 Let f be a function from the set A to the set B and let S be a subset of A. The *image* of S under the function f is the subset of B that consists of the images of the elements of S. We denote the image of S by $f(S)$, so

$$f(S) = \{ t \mid \exists s \in S\ (t = f(s)) \}.$$

We also use the shorthand $\{f(s) \mid s \in S\}$ to denote this set.

Remark: The notation $f(S)$ for the image of the set S under the function f is potentially ambiguous. Here, $f(S)$ denotes a set, and not the value of the function f for the set S.

EXAMPLE 7 Let $A = \{a, b, c, d, e\}$ and $B = \{1, 2, 3, 4\}$ with $f(a) = 2$, $f(b) = 1$, $f(c) = 4$, $f(d) = 1$, and $f(e) = 1$. The image of the subset $S = \{b, c, d\}$ is the set $f(S) = \{1, 4\}$. ◀

One-to-One and Onto Functions

Some functions never assign the same value to two different domain elements. These functions are said to be **one-to-one.**

DEFINITION 5 A function f is said to be *one-to-one,* or *injective,* if and only if $f(a) = f(b)$ implies that $a = b$ for all a and b in the domain of f. A function is said to be an *injection* if it is one-to-one.

Note that a function f is one-to-one if and only if $f(a) \neq f(b)$ whenever $a \neq b$. This way of expressing that f is one-to-one is obtained by taking the contrapositive of the implication in the definition.

Remark: We can express that f is one-to-one using quantifiers as $\forall a \forall b(f(a) = f(b) \rightarrow a = b)$ or equivalently $\forall a \forall b(a \neq b \rightarrow f(a) \neq f(b))$, where the universe of discourse is the domain of the function.

Assessment We illustrate this concept by giving examples of functions that are one-to-one and other functions that are not one-to-one.

EXAMPLE 8 Determine whether the function f from $\{a, b, c, d\}$ to $\{1, 2, 3, 4, 5\}$ with $f(a) = 4$, $f(b) = 5$, $f(c) = 1$, and $f(d) = 3$ is one-to-one.

Extra Examples *Solution:* The function f is one-to-one because f takes on different values at the four elements of its domain. This is illustrated in Figure 3. ◀

EXAMPLE 9 Determine whether the function $f(x) = x^2$ from the set of integers to the set of integers is one-to-one.

Solution: The function $f(x) = x^2$ is not one-to-one because, for instance, $f(1) = f(-1) = 1$, but $1 \neq -1$.

FIGURE 3 A One-to-One Function.

Note that the function $f(x) = x^2$ with its domain restricted to $\mathbf{Z}^+$ is one-to-one. (Technically, when we restrict the domain of a function, we obtain a new function whose values agree with those of the original function for the elements of the restricted domain. The restricted function is not defined for elements of the original domain outside of the restricted domain.) ◀

EXAMPLE 10 Determine whether the function $f(x) = x + 1$ from the set of real numbers to itself is one-to-one.

Solution: The function $f(x) = x + 1$ is a one-to-one function. To demonstrate this, note that $x + 1 \neq y + 1$ when $x \neq y$. ◀

We now give some conditions that guarantee that a function is one-to-one.

DEFINITION 6 A function f whose domain and codomain are subsets of the set of real numbers is called *increasing* if $f(x) \leq f(y)$, and *strictly increasing* if $f(x) < f(y)$, whenever $x < y$ and x and y are in the domain of f. Similarly, f is called *decreasing* if $f(x) \geq f(y)$, and *strictly decreasing* if $f(x) > f(y)$, whenever $x < y$ and x and y are in the domain of f. (The word *strictly* in this definition indicates a strict inequality.)

Remark: A function f is increasing if $\forall x \forall y(x < y \rightarrow f(x) \leq f(y))$, strictly increasing if $\forall x \forall y(x < y \rightarrow f(x) < f(y))$, decreasing if $\forall x \forall y(x < y \rightarrow f(x) \geq f(y))$, and strictly decreasing if $\forall x \forall y(x < y \rightarrow f(x) > f(y))$, where the universe of discourse is the domain of f.

From these definitions, we see that a function that is either strictly increasing or strictly decreasing must be one-to-one. However, a function that is increasing, but not strictly increasing, or decreasing, but not strictly decreasing, is not necessarily one-to-one.

For some functions the range and the codomain are equal. That is, every member of the codomain is the image of some element of the domain. Functions with this property are called **onto** functions.

DEFINITION 7 A function f from A to B is called *onto,* or *surjective,* if and only if for every element $b \in B$ there is an element $a \in A$ with $f(a) = b$. A function f is called a *surjection* if it is onto.

Remark: A function f is onto if $\forall y \exists x(f(x) = y)$, where the domain for x is the domain of the function and the domain for y is the codomain of the function.

We now give examples of onto functions and functions that are not onto.

FIGURE 4 An Onto Function.

EXAMPLE 11 Let f be the function from $\{a, b, c, d\}$ to $\{1, 2, 3\}$ defined by $f(a) = 3$, $f(b) = 2$, $f(c) = 1$, and $f(d) = 3$. Is f an onto function?

Extra Examples

Solution: Because all three elements of the codomain are images of elements in the domain, we see that f is onto. This is illustrated in Figure 4. Note that if the codomain were $\{1, 2, 3, 4\}$, then f would not be onto. ◀

EXAMPLE 12 Is the function $f(x) = x^2$ from the set of integers to the set of integers onto?

Solution: The function f is not onto because there is no integer x with $x^2 = -1$, for instance. ◀

EXAMPLE 13 Is the function $f(x) = x + 1$ from the set of integers to the set of integers onto?

Solution: This function is onto, because for every integer y there is an integer x such that $f(x) = y$. To see this, note that $f(x) = y$ if and only if $x + 1 = y$, which holds if and only if $x = y - 1$. ◀

DEFINITION 8 The function f is a *one-to-one correspondence,* or a *bijection,* if it is both one-to-one and onto.

Examples 14 and 15 illustrate the concept of a bijection.

EXAMPLE 14 Let f be the function from $\{a, b, c, d\}$ to $\{1, 2, 3, 4\}$ with $f(a) = 4$, $f(b) = 2$, $f(c) = 1$, and $f(d) = 3$. Is f a bijection?

Solution: The function f is one-to-one and onto. It is one-to-one because no two values in the domain are assigned the same function value. It is onto because all four elements of the codomain are images of elements in the domain. Hence, f is a bijection. ◀

Figure 5 displays four functions where the first is one-to-one but not onto, the second is onto but not one-to-one, the third is both one-to-one and onto, and the fourth is neither one-to-one nor onto. The fifth correspondence in Figure 5 is not a function, because it sends an element to two different elements.

Suppose that f is a function from a set A to itself. If A is finite, then f is one-to-one if and only if it is onto. (This follows from the result in Exercise 68 at the end of this section.) This is not necessarily the case if A is infinite (as will be shown in Section 2.4).

EXAMPLE 15 Let A be a set. The *identity function* on A is the function $\iota_A : A \to A$, where

$$\iota_A(x) = x$$

for all $x \in A$. In other words, the identity function ι_A is the function that assigns each element to itself. The function ι_A is one-to-one and onto, so it is a bijection. (Note that ι is the Greek letter iota.) ◀

FIGURE 5 Examples of Different Types of Correspondences.

Inverse Functions and Compositions of Functions

Now consider a one-to-one correspondence f from the set A to the set B. Because f is an onto function, every element of B is the image of some element in A. Furthermore, because f is also a one-to-one function, every element of B is the image of a *unique* element of A. Consequently, we can define a new function from B to A that reverses the correspondence given by f. This leads to Definition 9.

DEFINITION 9 Let f be a one-to-one correspondence from the set A to the set B. The *inverse function* of f is the function that assigns to an element b belonging to B the unique element a in A such that $f(a) = b$. The inverse function of f is denoted by f^{-1}. Hence, $f^{-1}(b) = a$ when $f(a) = b$.

Remark: Be sure not to confuse the function f^{-1} with the function $1/f$, which is the function that assigns to each x in the domain the value $1/f(x)$. Notice that the latter makes sense only when $f(x)$ is a non-zero real number. They are not the same.

Figure 6 illustrates the concept of an inverse function.

If a function f is not a one-to-one correspondence, we cannot define an inverse function of f. When f is not a one-to-one correspondence, either it is not one-to-one or it is not onto. If f is not one-to-one, some element b in the codomain is the image of more than one element in the domain. If f is not onto, for some element b in the codomain, no element a in the domain exists for which $f(a) = b$. Consequently, if f is not a one-to-one correspondence, we cannot assign to each element b in the codomain a unique element a in the domain such that $f(a) = b$ (because for some b there is either more than one such a or no such a).

FIGURE 6 The Function f^{-1} Is the Inverse of Function f.

A one-to-one correspondence is called **invertible** because we can define an inverse of this function. A function is **not invertible** if it is not a one-to-one correspondence, because the inverse of such a function does not exist.

EXAMPLE 16 Let f be the function from $\{a, b, c\}$ to $\{1, 2, 3\}$ such that $f(a) = 2$, $f(b) = 3$, and $f(c) = 1$. Is f invertible, and if it is, what is its inverse?

Solution: The function f is invertible because it is a one-to-one correspondence. The inverse function f^{-1} reverses the correspondence given by f, so $f^{-1}(1) = c$, $f^{-1}(2) = a$, and $f^{-1}(3) = b$. ◀

EXAMPLE 17 Let $f : \mathbf{Z} \to \mathbf{Z}$ be such that $f(x) = x + 1$. Is f invertible, and if it is, what is its inverse?

Solution: The function f has an inverse because it is a one-to-one correspondence, as we have shown. To reverse the correspondence, suppose that y is the image of x, so that $y = x + 1$. Then $x = y - 1$. This means that $y - 1$ is the unique element of $\mathbf{Z}$ that is sent to y by f. Consequently, $f^{-1}(y) = y - 1$. ◀

EXAMPLE 18 Let f be the function from $\mathbf{R}$ to $\mathbf{R}$ with $f(x) = x^2$. Is f invertible?

Solution: Because $f(-2) = f(2) = 4$, f is not one-to-one. If an inverse function were defined, it would have to assign two elements to 4. Hence, f is not invertible. ◀

Sometimes we can restrict the domain or the codomain of a function, or both, to obtain an invertible function, as Example 19 illustrates.

EXAMPLE 19 Show that if we restrict the function $f(x) = x^2$ in Example 18 to a function from the set of all nonnegative real numbers to the set of all nonnegative real numbers, then f is invertible.

Solution: The function $f(x) = x^2$ from the set of nonnegative real numbers to the set of nonnegative real numbers is one-to-one. To see this, note that if $f(x) = f(y)$, then $x^2 = y^2$, so $x^2 - y^2 = (x + y)(x - y) = 0$. This means that $x + y = 0$ or $x - y = 0$, so $x = -y$ or $x = y$. Because both x and y are nonnegative, we must have $x = y$. So, this function is one-to-one. Furthermore, $f(x) = x^2$ is onto when the codomain is the set of all nonnegative real numbers, because each nonnegative real number has a square root. That is, if y is a nonnegative real number, there exists a nonnegative real number x such that $x = \sqrt{y}$, which means that $x^2 = y$. Because the function $f(x) = x^2$ from the set of nonnegative real numbers to the set of nonnegative real numbers is one-to-one and onto, it is invertible. Its inverse is given by the rule $f^{-1}(y) = \sqrt{y}$. ◀

DEFINITION 10 Let g be a function from the set A to the set B and let f be a function from the set B to the set C. The *composition* of the functions f and g, denoted by $f \circ g$, is defined by

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

In other words, $f \circ g$ is the function that assigns to the element a of A the element assigned by f to $g(a)$. That is, to find $(f \circ g)(a)$ we first apply the function g to a to obtain $g(a)$ and then we apply the function f to the result $g(a)$ to obtain $(f \circ g)(a) = f(g(a))$. Note that the

FIGURE 7 The Composition of the Functions f and g.

composition $f \circ g$ cannot be defined unless the range of g is a subset of the domain of f. In Figure 7 the composition of functions is shown.

EXAMPLE 20 Let g be the function from the set $\{a, b, c\}$ to itself such that $g(a) = b$, $g(b) = c$, and $g(c) = a$. Let f be the function from the set $\{a, b, c\}$ to the set $\{1, 2, 3\}$ such that $f(a) = 3$, $f(b) = 2$, and $f(c) = 1$. What is the composition of f and g, and what is the composition of g and f?

Solution: The composition $f \circ g$ is defined by $(f \circ g)(a) = f(g(a)) = f(b) = 2$, $(f \circ g)(b) = f(g(b)) = f(c) = 1$, and $(f \circ g)(c) = f(g(c)) = f(a) = 3$.

Note that $g \circ f$ is not defined, because the range of f is not a subset of the domain of g. ◀

EXAMPLE 21 Let f and g be the functions from the set of integers to the set of integers defined by $f(x) = 2x + 3$ and $g(x) = 3x + 2$. What is the composition of f and g? What is the composition of g and f?

Solution: Both the compositions $f \circ g$ and $g \circ f$ are defined. Moreover,

$$(f \circ g)(x) = f(g(x)) = f(3x + 2) = 2(3x + 2) + 3 = 6x + 7$$

and

$$(g \circ f)(x) = g(f(x)) = g(2x + 3) = 3(2x + 3) + 2 = 6x + 11.$$ ◀

Remark: Note that even though $f \circ g$ and $g \circ f$ are defined for the functions f and g in Example 21, $f \circ g$ and $g \circ f$ are not equal. In other words, the commutative law does not hold for the composition of functions.

When the composition of a function and its inverse is formed, in either order, an identity function is obtained. To see this, suppose that f is a one-to-one correspondence from the set A to the set B. Then the inverse function f^{-1} exists and is a one-to-one correspondence from B to A. The inverse function reverses the correspondence of the original function, so $f^{-1}(b) = a$ when $f(a) = b$, and $f(a) = b$ when $f^{-1}(b) = a$. Hence,

$$(f^{-1} \circ f)(a) = f^{-1}(f(a)) = f^{-1}(b) = a,$$

and

$$(f \circ f^{-1})(b) = f(f^{-1}(b)) = f(a) = b.$$

FIGURE 8 The Graph of $f(n) = 2n + 1$ from Z to Z.

FIGURE 9 The Graph of $f(x) = x^2$ from Z to Z.

Consequently $f^{-1} \circ f = \iota_A$ and $f \circ f^{-1} = \iota_B$, where ι_A and ι_B are the identity functions on the sets A and B, respectively. That is, $(f^{-1})^{-1} = f$.

The Graphs of Functions

We can associate a set of pairs in $A \times B$ to each function from A to B. This set of pairs is called the **graph** of the function and is often displayed pictorially to aid in understanding the behavior of the function.

DEFINITION 11 Let f be a function from the set A to the set B. The *graph* of the function f is the set of ordered pairs $\{(a, b) \mid a \in A \text{ and } f(a) = b\}$.

From the definition, the graph of a function f from A to B is the subset of $A \times B$ containing the ordered pairs with the second entry equal to the element of B assigned by f to the first entry.

EXAMPLE 22 Display the graph of the function $f(n) = 2n + 1$ from the set of integers to the set of integers.

Solution: The graph of f is the set of ordered pairs of the form $(n, 2n + 1)$, where n is an integer. This graph is displayed in Figure 8. ◀

EXAMPLE 23 Display the graph of the function $f(x) = x^2$ from the set of integers to the set of integers.

Solution: The graph of f is the set of ordered pairs of the form $(x, f(x)) = (x, x^2)$, where x is an integer. This graph is displayed in Figure 9. ◀

Some Important Functions

Next, we introduce two important functions in discrete mathematics, namely, the floor and ceiling functions. Let x be a real number. The floor function rounds x down to the closest integer less than or equal to x, and the ceiling function rounds x up to the closest integer greater than or

(a) $y = \lfloor x \rfloor$ (b) $y = \lceil x \rceil$

FIGURE 10 Graphs of the (a) Floor and (b) Ceiling Functions.

equal to x. These functions are often used when objects are counted. They play an important role in the analysis of the number of steps used by procedures to solve problems of a particular size.

DEFINITION 12 The *floor function* assigns to the real number x the largest integer that is less than or equal to x. The value of the floor function at x is denoted by $\lfloor x \rfloor$. The *ceiling function* assigns to the real number x the smallest integer that is greater than or equal to x. The value of the ceiling function at x is denoted by $\lceil x \rceil$.

Remark: The floor function is often also called the *greatest integer function*. It is often denoted by $[x]$.

EXAMPLE 24 These are some values of the floor and ceiling functions:

$$\lfloor \tfrac{1}{2} \rfloor = 0,\ \lceil \tfrac{1}{2} \rceil = 1,\ \lfloor -\tfrac{1}{2} \rfloor = -1,\ \lceil -\tfrac{1}{2} \rceil = 0,\ \lfloor 3.1 \rfloor = 3,\ \lceil 3.1 \rceil = 4,\ \lfloor 7 \rfloor = 7,\ \lceil 7 \rceil = 7. \quad \blacktriangleleft$$

We display the graphs of the floor and ceiling functions in Figure 10. In Figure 10(a) we display the graph of the floor function $\lfloor x \rfloor$. Note that this function has the same value throughout the interval $[n, n+1)$, namely n, and then it jumps up to $n+1$ when $x = n+1$. In Figure 10(b) we display the graph of the ceiling function $\lceil x \rceil$. Note that this function has the same value throughout the interval $(n, n+1]$, namely $n+1$, and then jumps to $n+2$ when x is a little larger than $n+1$.

Links

The floor and ceiling functions are useful in a wide variety of applications, including those involving data storage and data transmission. Consider Examples 25 and 26, typical of basic calculations done when database and data communications problems are studied.

EXAMPLE 25 Data stored on a computer disk or transmitted over a data network are usually represented as a string of bytes. Each byte is made up of 8 bits. How many bytes are required to encode 100 bits of data?

Solution: To determine the number of bytes needed, we determine the smallest integer that is at least as large as the quotient when 100 is divided by 8, the number of bits in a byte. Consequently, $\lceil 100/8 \rceil = \lceil 12.5 \rceil = 13$ bytes are required. ◀

TABLE 1 Useful Properties of the Floor and Ceiling Functions. (n is an integer)
(1a) $\lfloor x \rfloor = n$ if and only if $n \leq x < n+1$ (1b) $\lceil x \rceil = n$ if and only if $n-1 < x \leq n$ (1c) $\lfloor x \rfloor = n$ if and only if $x-1 < n \leq x$ (1d) $\lceil x \rceil = n$ if and only if $x \leq n < x+1$
(2) $x - 1 < \lfloor x \rfloor \leq x \leq \lceil x \rceil < x + 1$
(3a) $\lfloor -x \rfloor = -\lceil x \rceil$ (3b) $\lceil -x \rceil = -\lfloor x \rfloor$
(4a) $\lfloor x + n \rfloor = \lfloor x \rfloor + n$ (4b) $\lceil x + n \rceil = \lceil x \rceil + n$

EXAMPLE 26 In asynchronous transfer mode (ATM) (a communications protocol used on backbone networks), data are organized into cells of 53 bytes. How many ATM cells can be transmitted in 1 minute over a connection that transmits data at the rate of 500 kilobits per second?

Solution: In 1 minute, this connection can transmit $500{,}000 \cdot 60 = 30{,}000{,}000$ bits. Each ATM cell is 53 bytes long, which means that it is $53 \cdot 8 = 424$ bits long. To determine the number of cells that can be transmitted in 1 minute, we determine the largest integer not exceeding the quotient when 30,000,000 is divided by 424. Consequently, $\lfloor 30{,}000{,}000/424 \rfloor = 70{,}754$ ATM cells can be transmitted in 1 minute over a 500 kilobit per second connection. ◀

Table 1, with x denoting a real number, displays some simple but important properties of the floor and ceiling functions. Because these functions appear so frequently in discrete mathematics, it is useful to look over these identities. Each property in this table can be established using the definitions of the floor and ceiling functions. Properties (1a), (1b), (1c), and (1d) follow directly from these definitions. For example, (1a) states that $\lfloor x \rfloor = n$ if and only if the integer n is less than or equal to x and $n+1$ is larger than x. This is precisely what it means for n to be the greatest integer not exceeding x, which is the definition of $\lfloor x \rfloor = n$. Properties (1b), (1c), and (1d) can be established similarly. We will prove property (4a) using a direct proof.

Proof: Suppose that $\lfloor x \rfloor = m$, where m is a positive integer. By property (1a), it follows that $m \leq x < m+1$. Adding n to both sides of this inequality shows that $m + n \leq x + n < m + n + 1$. Using property (1a) again, we see that $\lfloor x + n \rfloor = m + n = \lfloor x \rfloor + n$. This completes the proof. Proofs of the other properties are left as exercises. ◁

The floor and ceiling functions enjoy many other useful properties besides those displayed in Table 1. There are also many statements about these functions that may appear to be correct, but actually are not. We will consider statements about the floor and ceiling functions in Examples 27 and 28.

A useful approach for considering statements about the floor function is to let $x = n + \epsilon$, where $n = \lfloor x \rfloor$ is an integer, and ϵ, the fractional part of x, satisfies the inequality $0 \leq \epsilon < 1$. Similarly, when considering statements about the ceiling function, it is useful to write $x = n - \epsilon$, where $n = \lceil x \rceil$ is an integer and $0 \leq \epsilon < 1$.

EXAMPLE 27 Prove that if x is a real number, then $\lfloor 2x \rfloor = \lfloor x \rfloor + \lfloor x + \frac{1}{2} \rfloor$.

Extra Examples

Solution: To prove this statement we let $x = n + \epsilon$, where n is a positive integer and $0 \leq \epsilon < 1$. There are two cases to consider, depending on whether ϵ is less than or greater than or equal to $\frac{1}{2}$. (The reason we choose these two cases will be made clear in the proof.)

We first consider the case when $0 \leq \epsilon < \frac{1}{2}$. In this case, $2x = 2n + 2\epsilon$ and $\lfloor 2x \rfloor = 2n$ because $0 \leq 2\epsilon < 1$. Similarly, $x + \frac{1}{2} = n + (\frac{1}{2} + \epsilon)$, so $\lfloor x + \frac{1}{2} \rfloor = n$, because $0 < \frac{1}{2} + \epsilon < 1$. Consequently, $\lfloor 2x \rfloor = 2n$ and $\lfloor x \rfloor + \lfloor x + \frac{1}{2} \rfloor = n + n = 2n$.

Next, we consider the case when $\frac{1}{2} \leq \epsilon < 1$. In this case, $2x = 2n + 2\epsilon = (2n + 1) + (2\epsilon - 1)$. Because $0 \leq 2\epsilon - 1 < 1$, it follows that $\lfloor 2x \rfloor = 2n + 1$. Because $\lfloor x + \frac{1}{2} \rfloor = \lfloor n + (\frac{1}{2} + \epsilon) \rfloor = \lfloor n + 1 + (\epsilon - \frac{1}{2}) \rfloor$ and $0 \leq \epsilon - \frac{1}{2} < 1$, it follows that $\lfloor x + \frac{1}{2} \rfloor = n + 1$. Consequently, $\lfloor 2x \rfloor = 2n + 1$ and $\lfloor x \rfloor + \lfloor x + \frac{1}{2} \rfloor = n + (n + 1) = 2n + 1$. This concludes the proof. ◀

EXAMPLE 28 Prove or disprove that $\lceil x + y \rceil = \lceil x \rceil + \lceil y \rceil$ for all real numbers x and y.

Solution: Although this statement may appear reasonable, it is false. A counterexample is supplied by $x = \frac{1}{2}$ and $y = \frac{1}{2}$. With these values we find that $\lceil x + y \rceil = \lceil \frac{1}{2} + \frac{1}{2} \rceil = \lceil 1 \rceil = 1$, but $\lceil x \rceil + \lceil y \rceil = \lceil \frac{1}{2} \rceil + \lceil \frac{1}{2} \rceil = 1 + 1 = 2$. ◀

There are certain types of functions that will be used throughout the text. These include polynomial, logarithmic, and exponential functions. A brief review of the properties of these functions needed in this text is given in Appendix 2. In this book the notation $\log x$ will be used to denote the logarithm to the base 2 of x, because 2 is the base that we will usually use for logarithms. We will denote logarithms to the base b, where b is any real number greater than 1, by $\log_b x$, and the natural logarithm by $\ln x$.

Another function we will use throughout this text is the **factorial function** $f : \mathbf{N} \to \mathbf{Z}^+$, denoted by $f(n) = n!$. The value of $f(n) = n!$ is the product of the first n positive integers, so $f(n) = 1 \cdot 2 \cdots (n - 1) \cdot n$ [and $f(0) = 0! = 1$].

EXAMPLE 29 We have $f(1) = 1! = 1$, $f(2) = 2! = 1 \cdot 2 = 2$, $f(6) = 6! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 = 720$, and $f(20) = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12 \cdot 13 \cdot 14 \cdot 15 \cdot 16 \cdot 17 \cdot 18 \cdot 19 \cdot 20 = 2{,}432{,}902{,}008{,}176{,}640{,}000$. ◀

Links

JAMES STIRLING (1692–1770) was born near the town of Stirling, Scotland. His family strongly supported the Jacobite cause of the Stuarts as an alternative to the British crown. The first information known about James is that he entered Balliol College, Oxford, on a scholarship in 1711. However, he later lost his scholarship when he refused to pledge his allegiance to the British crown. The first Jacobean rebellion took place in 1715, and Stirling was accused of communicating with rebels. He was charged with cursing King George, but he was acquitted of these charges. Even though he could not graduate from Oxford because of his politics, he remained there for several years. Stirling published his first work, which extended Newton's work on plane curves, in 1717. He traveled to Venice, where a chair of mathematics had been promised to him, an appointment that unfortunately fell through. Nevertheless, Stirling stayed in Venice, continuing his mathematical work. He attended the University of Padua in 1721, and in 1722 he returned to Glasgow. Stirling apparently fled Italy after learning the secrets of the Italian glass industry, avoiding the efforts of Italian glass makers to assassinate him to protect their secrets.

In late 1724 Stirling moved to London, staying there 10 years teaching mathematics and actively engaging in research. In 1730 he published *Methodus Differentialis*, his most important work, presenting results on infinite series, summations, interpolation, and quadrature. It is in this book that his asymptotic formula for $n!$ appears. Stirling also worked on gravitation and the shape of the earth; he stated, but did not prove, that the earth is an oblate spheroid. Stirling returned to Scotland in 1735, when he was appointed manager of a Scottish mining company. He was very successful in this role and even published a paper on the ventilation of mine shafts. He continued his mathematical research, but at a reduced pace, during his years in the mining industry. Stirling is also noted for surveying the River Clyde with the goal of creating a series of locks to make it navigable. In 1752 the citizens of Glasgow presented him with a silver teakettle as a reward for this work.

Example 29 illustrates that the factorial function grows extremely rapidly as n grows. The rapid growth of the factorial function is made clearer by Stirling's formula, a result from higher mathematics that tell us that $n! \sim \sqrt{2\pi n}(n/e)^n$. Here, we have used the notation $f(n) \sim g(n)$, which means that the ratio $f(n)/g(n)$ approaches 1 as n grows without bound (that is, $\lim_{n\to\infty} f(n)/g(n) = 1$). The symbol $\sim$ is read "is asymptotic to." Stirling's formula is named after James Stirling, a Scottish mathematician of the eighteenth century.

Exercises

1. Why is f not a function from $\mathbf{R}$ to $\mathbf{R}$ if
   a) $f(x) = 1/x$?
   b) $f(x) = \sqrt{x}$?
   c) $f(x) = \pm\sqrt{(x^2+1)}$?
2. Determine whether f is a function from $\mathbf{Z}$ to $\mathbf{R}$ if
   a) $f(n) = \pm n$.
   b) $f(n) = \sqrt{n^2+1}$.
   c) $f(n) = 1/(n^2-4)$.
3. Determine whether f is a function from the set of all bit strings to the set of integers if
   a) $f(S)$ is the position of a 0 bit in S.
   b) $f(S)$ is the number of 1 bits in S.
   c) $f(S)$ is the smallest integer i such that the ith bit of S is 1 and $f(S) = 0$ when S is the empty string, the string with no bits.
4. Find the domain and range of these functions. Note that in each case, to find the domain, determine the set of elements assigned values by the function.
   a) the function that assigns to each nonnegative integer its last digit
   b) the function that assigns the next largest integer to a positive integer
   c) the function that assigns to a bit string the number of one bits in the string
   d) the function that assigns to a bit string the number of bits in the string
5. Find the domain and range of these functions. Note that in each case, to find the domain, determine the set of elements assigned values by the function.
   a) the function that assigns to each bit string the number of ones minus the number of zeros
   b) the function that assigns to each bit string twice the number of zeros in that string
   c) the function that assigns the number of bits left over when a bit string is split into bytes (which are blocks of 8 bits)
   d) the function that assigns to each positive integer the largest perfect square not exceeding this integer
6. Find the domain and range of these functions.
   a) the function that assigns to each pair of positive integers the first integer of the pair
   b) the function that assigns to each positive integer its largest decimal digit
   c) the function that assigns to a bit string the number of ones minus the number of zeros in the string
   d) the function that assigns to each positive integer the largest integer not exceeding the square root of the integer
   e) the function that assigns to a bit string the longest string of ones in the string
7. Find the domain and range of these functions.
   a) the function that assigns to each pair of positive integers the maximum of these two integers
   b) the function that assigns to each positive integer the number of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 that do not appear as decimal digits of the integer
   c) the function that assigns to a bit string the number of times the block 11 appears
   d) the function that assigns to a bit string the numerical position of the first 1 in the string and that assigns the value 0 to a bit string consisting of all 0s
8. Find these values.
   a) $\lfloor 1.1 \rfloor$ b) $\lceil 1.1 \rceil$
   c) $\lfloor -0.1 \rfloor$ d) $\lceil -0.1 \rceil$
   e) $\lceil 2.99 \rceil$ f) $\lceil -2.99 \rceil$
   g) $\lfloor \frac{1}{2} + \lceil \frac{1}{2} \rceil \rfloor$ h) $\lceil \lfloor \frac{1}{2} \rfloor + \lceil \frac{1}{2} \rceil + \frac{1}{2} \rceil$
9. Find these values.
   a) $\lceil \frac{3}{4} \rceil$ b) $\lfloor \frac{7}{8} \rfloor$
   c) $\lceil -\frac{3}{4} \rceil$ d) $\lfloor -\frac{7}{8} \rfloor$
   e) $\lceil 3 \rceil$ f) $\lfloor -1 \rfloor$
   g) $\lfloor \frac{1}{2} + \lceil \frac{3}{2} \rceil \rfloor$ h) $\lfloor \frac{1}{2} \cdot \lfloor \frac{5}{2} \rfloor \rfloor$
10. Determine whether each of these functions from $\{a, b, c, d\}$ to itself is one-to-one.
    a) $f(a) = b, f(b) = a, f(c) = c, f(d) = d$
    b) $f(a) = b, f(b) = b, f(c) = d, f(d) = c$
    c) $f(a) = d, f(b) = b, f(c) = c, f(d) = d$
11. Which functions in Exercise 10 are onto?
12. Determine whether each of these functions from $\mathbf{Z}$ to $\mathbf{Z}$ is one-to-one.
    a) $f(n) = n - 1$ b) $f(n) = n^2 + 1$
    c) $f(n) = n^3$ d) $f(n) = \lceil n/2 \rceil$
13. Which functions in Exercise 12 are onto?
14. Determine whether $f\colon \mathbf{Z} \times \mathbf{Z} \to \mathbf{Z}$ is onto if
    a) $f(m, n) = 2m - n$.
    b) $f(m, n) = m^2 - n^2$.

c) $f(m, n) = m + n + 1$.
d) $f(m, n) = |m| - |n|$.
e) $f(m, n) = m^2 - 4$.

15. Determine whether the function $f: \mathbf{Z} \times \mathbf{Z} \to \mathbf{Z}$ is onto if
a) $f(m, n) = m + n$.
b) $f(m, n) = m^2 + n^2$.
c) $f(m, n) = m$.
d) $f(m, n) = |n|$.
e) $f(m, n) = m - n$.

16. Give an example of a function from $\mathbf{N}$ to $\mathbf{N}$ that is
a) one-to-one but not onto.
b) onto but not one-to-one.
c) both onto and one-to-one (but different from the identity function).
d) neither one-to-one nor onto.

17. Give an explicit formula for a function from the set of integers to the set of positive integers that is
a) one-to-one, but not onto.
b) onto, but not one-to-one.
c) one-to-one and onto.
d) neither one-to-one nor onto.

18. Determine whether each of these functions is a bijection from $\mathbf{R}$ to $\mathbf{R}$.
a) $f(x) = -3x + 4$
b) $f(x) = -3x^2 + 7$
c) $f(x) = (x + 1)/(x + 2)$
d) $f(x) = x^5 + 1$

19. Determine whether each of these functions is a bijection from $\mathbf{R}$ to $\mathbf{R}$.
a) $f(x) = 2x + 1$
b) $f(x) = x^2 + 1$
c) $f(x) = x^3$
d) $f(x) = (x^2 + 1)/(x^2 + 2)$

20. Let $f: \mathbf{R} \to \mathbf{R}$ and let $f(x) > 0$ for all $x \in \mathbf{R}$. Show that $f(x)$ is strictly increasing if and only if the function $g(x) = 1/f(x)$ is strictly decreasing.

21. Let $f: \mathbf{R} \to \mathbf{R}$ and let $f(x) > 0$. Show that $f(x)$ is strictly decreasing if and only if the function $g(x) = 1/f(x)$ is strictly increasing.

22. Give an example of an increasing function with the set of real numbers as its domain and codomain that is not one-to-one.

23. Give an example of a decreasing function with the set of real numbers as its domain and codomain that is not one-to-one.

24. Show that the function $f(x) = e^x$ from the set of real number to the set of real numbers is not invertible, but if the codomain is restricted to the set of positive real numbers, the resulting function is invertible.

25. Show that the function $f(x) = |x|$ from the set of real numbers to the set of nonnegative real numbers is not invertible, but if the domain is restricted to the set of nonnegative real numbers, the resulting function is invertible.

26. Let $S = \{-1, 0, 2, 4, 7\}$. Find $f(S)$ if
a) $f(x) = 1$.
b) $f(x) = 2x + 1$.
c) $f(x) = \lceil x/5 \rceil$.
d) $f(x) = \lfloor (x^2 + 1)/3 \rfloor$.

27. Let $f(x) = \lfloor x^2/3 \rfloor$. Find $f(S)$ if
a) $S = \{-2, -1, 0, 1, 2, 3\}$.
b) $S = \{0, 1, 2, 3, 4, 5\}$.
c) $S = \{1, 5, 7, 11\}$.
d) $S = \{2, 6, 10, 14\}$.

28. Let $f(x) = 2x$. What is
a) $f(\mathbf{Z})$?
b) $f(\mathbf{N})$?
c) $f(\mathbf{R})$?

29. Suppose that g is a function from A to B and f is a function from B to C.
a) Show that if both f and g are one-to-one functions, then $f \circ g$ is also one-to-one.
b) Show that if both f and g are onto functions, then $f \circ g$ is also onto.

***30.** If f and $f \circ g$ are one-to-one, does it follow that g is one-to-one? Justify your answer.

***31.** If f and $f \circ g$ are onto, does it follow that g is onto? Justify your answer.

32. Find $f \circ g$ and $g \circ f$, where $f(x) = x^2 + 1$ and $g(x) = x + 2$, are functions from $\mathbf{R}$ to $\mathbf{R}$.

33. Find $f + g$ and fg for the functions f and g given in Exercise 32.

34. Let $f(x) = ax + b$ and $g(x) = cx + d$, where a, b, c, and d are constants. Determine for which constants a, b, c, and d it is true that $f \circ g = g \circ f$.

35. Show that the function $f(x) = ax + b$ from $\mathbf{R}$ to $\mathbf{R}$ is invertible, where a and b are constants, with $a \neq 0$, and find the inverse of f.

36. Let f be a function from the set A to the set B. Let S and T be subsets of A. Show that
a) $f(S \cup T) = f(S) \cup f(T)$.
b) $f(S \cap T) \subseteq f(S) \cap f(T)$.

37. Give an example to show that the inclusion in part (b) in Exercise 36 may be proper.

Let f be a function from the set A to the set B. Let S be a subset of B. We define the **inverse image** of S to be the subset of A whose elements are precisely all pre-images of all elements of S. We denote the inverse image of S by $f^{-1}(S)$, so $f^{-1}(S) = \{a \in A \mid f(a) \in S\}$. (*Beware:* The notation f^{-1} is used in two different ways. Do not confuse the notation introduced here with the notation $f^{-1}(y)$ for the value at y of the inverse of the invertible function f. Notice also that $f^{-1}(S)$, the inverse image of the set S, makes sense for all functions f, not just invertible functions.)

38. Let f be the function from $\mathbf{R}$ to $\mathbf{R}$ defined by $f(x) = x^2$. Find
a) $f^{-1}(\{1\})$.
b) $f^{-1}(\{x \mid 0 < x < 1\})$.
c) $f^{-1}(\{x \mid x > 4\})$.

39. Let $g(x) = \lfloor x \rfloor$. Find
a) $g^{-1}(\{0\})$.
b) $g^{-1}(\{-1, 0, 1\})$.
c) $g^{-1}(\{x \mid 0 < x < 1\})$.

40. Let f be a function from A to B. Let S and T be subsets of B. Show that

a) $f^{-1}(S \cup T) = f^{-1}(S) \cup f^{-1}(T)$.
b) $f^{-1}(S \cap T) = f^{-1}(S) \cap f^{-1}(T)$.

41. Let f be a function from A to B. Let S be a subset of B. Show that $f^{-1}(\overline{S}) = \overline{f^{-1}(S)}$.

42. Show that $\lfloor x + \frac{1}{2} \rfloor$ is the closest integer to the number x, except when x is midway between two integers, when it is the larger of these two integers.

43. Show that $\lceil x - \frac{1}{2} \rceil$ is the closest integer to the number x, except when x is midway between two integers, when it is the smaller of these two integers.

44. Show that if x is a real number, then $\lceil x \rceil - \lfloor x \rfloor = 1$ if x is not an integer and $\lceil x \rceil - \lfloor x \rfloor = 0$ if x is an integer.

45. Show that if x is a real number, then $x - 1 < \lfloor x \rfloor \leq x \leq \lceil x \rceil < x + 1$.

46. Show that if x is a real number and m is an integer, then $\lceil x + m \rceil = \lceil x \rceil + m$.

47. Show that if x is a real number and n is an integer, then

a) $x < n$ if and only if $\lfloor x \rfloor < n$.
b) $n < x$ if and only if $n < \lceil x \rceil$.

48. Show that if x is a real number and n is an integer, then

a) $x \leq n$ if and only if $\lceil x \rceil \leq n$.
b) $n \leq x$ if and only if $n \leq \lfloor x \rfloor$.

49. Prove that if n is an integer, then $\lfloor n/2 \rfloor = n/2$ if n is even and $(n-1)/2$ if n is odd.

50. Prove that if x is a real number, then $\lfloor -x \rfloor = -\lceil x \rceil$ and $\lceil -x \rceil = -\lfloor x \rfloor$.

51. The function INT is found on some calculators, where $\text{INT}(x) = \lfloor x \rfloor$ when x is a nonnegative real number and $\text{INT}(x) = \lceil x \rceil$ when x is a negative real number. Show that this INT function satisfies the identity $\text{INT}(-x) = -\text{INT}(x)$.

52. Let a and b be real numbers with $a < b$. Use the floor and/or ceiling functions to express the number of integers n that satisfy the inequality $a \leq n \leq b$.

53. Let a and b be real numbers with $a < b$. Use the floor and/or ceiling functions to express the number of integers n that satisfy the inequality $a < n < b$.

54. How many bytes are required to encode n bits of data where n equals

a) 4? **b)** 10? **c)** 500? **d)** 3000?

55. How many bytes are required to encode n bits of data where n equals

a) 7? **b)** 17? **c)** 1001? **d)** 28,800?

56. How many ATM cells (described in Example 26) can be transmitted in 10 seconds over a link operating at the following rates?

a) 128 kilobits per second (1 kilobit = 1000 bits)
b) 300 kilobits per second
c) 1 megabit per second (1 megabit = 1,000,000 bits)

57. Data are transmitted over a particular Ethernet network in blocks of 1500 octets (blocks of 8 bits). How many blocks are required to transmit the following amounts of data over this Ethernet network? (Note that a byte is a synonym for an octet, a kilobyte is 1000 bytes, and a megabyte is 1,000,000 bytes.)

a) 150 kilobytes of data
b) 384 kilobytes of data
c) 1.544 megabytes of data
d) 45.3 megabytes of data

58. Draw the graph of the function $f(n) = 1 - n^2$ from $\mathbf{Z}$ to $\mathbf{Z}$.

59. Draw the graph of the function $f(x) = \lfloor 2x \rfloor$ from $\mathbf{R}$ to $\mathbf{R}$.

60. Draw the graph of the function $f(x) = \lfloor x/2 \rfloor$ from $\mathbf{R}$ to $\mathbf{R}$.

61. Draw the graph of the function $f(x) = \lfloor x \rfloor + \lfloor x/2 \rfloor$ from $\mathbf{R}$ to $\mathbf{R}$.

62. Draw the graph of the function $f(x) = \lceil x \rceil + \lfloor x/2 \rfloor$ from $\mathbf{R}$ to $\mathbf{R}$.

63. Draw graphs of each of these functions.

a) $f(x) = \lfloor x + \frac{1}{2} \rfloor$ **b)** $f(x) = \lfloor 2x + 1 \rfloor$
c) $f(x) = \lceil x/3 \rceil$ **d)** $f(x) = \lceil 1/x \rceil$
e) $f(x) = \lceil x - 2 \rceil + \lfloor x + 2 \rfloor$
f) $f(x) = \lfloor 2x \rfloor \lceil x/2 \rceil$ **g)** $f(x) = \lceil \lfloor x - \frac{1}{2} \rfloor + \frac{1}{2} \rceil$

64. Draw graphs of each of these functions.

a) $f(x) = \lceil 3x - 2 \rceil$ **b)** $f(x) = \lceil 0.2x \rceil$
c) $f(x) = \lfloor -1/x \rfloor$ **d)** $f(x) = \lfloor x^2 \rfloor$
e) $f(x) = \lceil x/2 \rceil \lfloor x/2 \rfloor$ **f)** $f(x) = \lfloor x/2 \rfloor + \lceil x/2 \rfloor$
g) $f(x) = \lfloor 2 \lceil x/2 \rceil + \frac{1}{2} \rfloor$

65. Find the inverse function of $f(x) = x^3 + 1$.

66. Suppose that f is an invertible function from Y to Z and g is an invertible function from X to Y. Show that the inverse of the composition $f \circ g$ is given by $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$.

67. Let S be a subset of a universal set U. The **characteristic function** f_S of S is the function from U to the set $\{0, 1\}$ such that $f_S(x) = 1$ if x belongs to S and $f_S(x) = 0$ if x does not belong to S. Let A and B be sets. Show that for all x,

a) $f_{A \cap B}(x) = f_A(x) \cdot f_B(x)$
b) $f_{A \cup B}(x) = f_A(x) + f_B(x) - f_A(x) \cdot f_B(x)$
c) $f_{\overline{A}}(x) = 1 - f_A(x)$
d) $f_{A \oplus B}(x) = f_A(x) + f_B(x) - 2 f_A(x) f_B(x)$

68. Suppose that f is a function from A to B, where A and B are finite sets with $|A| = |B|$. Show that f is one-to-one if and only if it is onto.

69. Prove or disprove each of these statements about the floor and ceiling functions.

a) $\lceil \lfloor x \rfloor \rceil = \lfloor x \rfloor$ for all real numbers x.
b) $\lfloor 2x \rfloor = 2 \lfloor x \rfloor$ whenever x is a real number.
c) $\lceil x \rceil + \lceil y \rceil - \lceil x + y \rceil = 0$ or 1 whenever x and y are real numbers.

d) $\lceil xy \rceil = \lceil x \rceil \lceil y \rceil$ for all real numbers x and y.

e) $\left\lceil \frac{x}{2} \right\rceil = \left\lfloor \frac{x+1}{2} \right\rfloor$ for all real numbers x.

70. Prove or disprove each of these statements about the floor and ceiling functions.

a) $\lfloor \lceil x \rceil \rfloor = \lceil x \rceil$ for all real numbers x.

b) $\lfloor x + y \rfloor = \lfloor x \rfloor + \lfloor y \rfloor$ for all real numbers x and y.

c) $\lceil \lceil x/2 \rceil /2 \rceil = \lceil x/4 \rceil$ for all real numbers x.

d) $\lfloor \sqrt{\lceil x \rceil} \rfloor = \lfloor \sqrt{x} \rfloor$ for all positive real numbers x.

e) $\lfloor x \rfloor + \lfloor y \rfloor + \lfloor x + y \rfloor \leq \lfloor 2x \rfloor + \lfloor 2y \rfloor$ for all real numbers x and y.

71. Prove that if x is a positive real number, then

a) $\lfloor \sqrt{\lfloor x \rfloor} \rfloor = \lfloor \sqrt{x} \rfloor$.

b) $\lceil \sqrt{\lceil x \rceil} \rceil = \lceil \sqrt{x} \rceil$.

72. Let x be a real number. Show that $\lfloor 3x \rfloor = \lfloor x \rfloor + \lfloor x + \frac{1}{3} \rfloor + \lfloor x + \frac{2}{3} \rfloor$.

A program designed to evaluate a function may not produce the correct value of the function for all elements in the domain of this function. For example, a program may not produce a correct value because evaluating the function may lead to an infinite loop or an overflow. Similarly, in abstract mathematics, we often want to discuss functions that are defined only for a subset of the real numbers, such as $1/x$, $\sqrt{x}$, and $\arcsin(x)$. Also, we may want to use such notions as the "youngest child" function, which is undefined for a couple having no children, or the "time of sunrise," which is undefined for some days above the Arctic Circle.

To study such situations, we use the concept of a partial function. A **partial function** f from a set A to a set B is an assignment to each element a in a subset of A, called the **domain of definition** of f, of a unique element b in B. The sets A and B are called the **domain** and **codomain** of f, respectively. We say that f is **undefined** for elements in A that are not in the domain of definition of f. We write $f : A \to B$ to denote that f is a partial function from A to B. (This is the same notation as is used for functions. The context in which the notation is used determines whether f is a partial function or a total function.) When the domain of definition of f equals A, we say that f is a **total function.**

73. For each of these partial functions, determine its domain, codomain, domain of definition, and the set of values for which it is undefined. Also, determine whether it is a total function.

a) $f : \mathbf{Z} \to \mathbf{R}$, $f(n) = 1/n$

b) $f : \mathbf{Z} \to \mathbf{Z}$, $f(n) = \lceil n/2 \rceil$

c) $f : \mathbf{Z} \times \mathbf{Z} \to \mathbf{Q}$, $f(m, n) = m/n$

d) $f : \mathbf{Z} \times \mathbf{Z} \to \mathbf{Z}$, $f(m, n) = mn$

e) $f : \mathbf{Z} \times \mathbf{Z} \to \mathbf{Z}$, $f(m, n) = m - n$ if $m > n$

74. a) Show that a partial function from A to B can be viewed as a function f^* from A to $B \cup \{u\}$, where u is not an element of B and

$$f^*(a) = \begin{cases} f(a) & \text{if } a \text{ belongs to the domain of definition of } f \\ u & \text{if } f \text{ is undefined at } a. \end{cases}$$

b) Using the construction in (a), find the function f^* corresponding to each partial function in Exercise 73.

☞**75. a)** Show that if a set S has cardinality m, where m is a positive integer, then there is a one-to-one correspondence between S and the set $\{1, 2, \ldots, m\}$.

b) Show that if S and T are two sets each with m elements, where m is a positive integer, then there is a one-to-one correspondence between S and T.

***76.** Show that a set S is infinite if and only if there is a proper subset A of S such that there is a one-to-one correspondence between A and S.

***77.** Show that the polynomial function $f : \mathbf{Z}^+ \times \mathbf{Z}^+ \to \mathbf{Z}^+$ with $f(m, n) = (m + n - 2)(m + n - 1)/2 + m$ is one-to-one and onto.

***78.** Show that when you substitute $(3n + 1)^2$ for each occurrence of n and $(3m + 1)^2$ for each occurrence of m in the right-hand side of the formula for the function $f(m, n)$ in Exercise 77, you obtain a one-to-one polynomial function $\mathbf{Z} \times \mathbf{Z} \to \mathbf{Z}$. It is an open question whether there is a one-to-one polynomial function $\mathbf{Q} \times \mathbf{Q} \to \mathbf{Q}$.

2.4 Sequences and Summations

Introduction

Sequences are ordered lists of elements. Sequences are used in discrete mathematics in many ways. They can be used to represent solutions to certain counting problems, as we will see in Chapter 7. They are also an important data structure in computer science. This section contains a review of the notation used to represent sequences and sums of terms of sequences.

When the elements of an infinite set can be listed, the set is called countable. We will conclude this section with a discussion of both countable and uncountable sets. We will prove that the set of rational numbers is countable, but the set of real numbers is not.

Sequences

A sequence is a discrete structure used to represent an ordered list. For example, 1, 2, 3, 5, 8 is a sequence with five terms and 1, 3, 9, 27, 81, ..., 30, ... is an infinite sequence.

DEFINITION 1 A *sequence* is a function from a subset of the set of integers (usually either the set $\{0, 1, 2, \ldots\}$ or the set $\{1, 2, 3, \ldots\}$) to a set S. We use the notation a_n to denote the image of the integer n. We call a_n a *term* of the sequence.

We use the notation $\{a_n\}$ to describe the sequence. (Note that a_n represents an individual term of the sequence $\{a_n\}$. Also note that the notation $\{a_n\}$ for a sequence conflicts with the notation for a set. However, the context in which we use this notation will always make it clear when we are dealing with sets and when we are dealing with sequences. Note that although we have used the letter a in the notation for a sequence, other letters or expressions may be used depending on the sequence under consideration. That is, the choice of the letter a is arbitrary.)

We describe sequences by listing the terms of the sequence in order of increasing subscripts.

EXAMPLE 1 Consider the sequence $\{a_n\}$, where

$$a_n = \frac{1}{n}.$$

The list of the terms of this sequence, beginning with a_1, namely,

$$a_1, a_2, a_3, a_4, \ldots,$$

starts with

$$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots.$$ ◀

DEFINITION 2 A *geometric progression* is a sequence of the form

$$a, ar, ar^2, \ldots, ar^n, \ldots$$

where the *initial term* a and the *common ratio* r are real numbers.

Remark: A geometric progression is a discrete analogue of the exponential function $f(x) = ar^x$.

EXAMPLE 2 The sequences $\{b_n\}$ with $b_n = (-1)^n$, $\{c_n\}$ with $c_n = 2 \cdot 5^n$, and $\{d_n\}$ with $d_n = 6 \cdot (1/3)^n$ are geometric progressions with initial term and common ratio equal to 1 and -1; 2 and 5; and 6 and $1/3$, respectively, if we start at $n = 0$. The list of terms $b_0, b_1, b_2, b_3, b_4, \ldots$ begins with

$$1, -1, 1, -1, 1, \ldots;$$

the list of terms $c_0, c_1, c_2, c_3, c_4, \ldots$ begins with

$$2, 10, 50, 250, 1250, \ldots;$$

and the list of terms $d_0, d_1, d_2, d_3, d_4, \ldots$ begins with

$$6, 2, \frac{2}{3}, \frac{2}{9}, \frac{2}{27}, \ldots.$$ ◀

DEFINITION 3 An *arithmetic progression* is a sequence of the form

$$a, a+d, a+2d, \ldots, a+nd, \ldots$$

where the *initial term* a and the *common difference* d are real numbers.

Remark: An arithmetic progression is a discrete analogue of the linear function $f(x) = dx + a$.

EXAMPLE 3 The sequences $\{s_n\}$ with $s_n = -1 + 4n$ and $\{t_n\}$ with $t_n = 7 - 3n$ are both arithmetic progressions with initial terms and common differences equal to -1 and 4, and 7 and -3, respectively, if we start at $n = 0$. The list of terms $s_0, s_1, s_2, s_3, \ldots$ begins with

$$-1, 3, 7, 11, \ldots,$$

and the list of terms $t_0, t_1, t_2, t_3, \ldots$ begins with

$$7, 4, 1, -2, \ldots.$$ ◀

Sequences of the form $a_1, a_2, \ldots, a_n$ are often used in computer science. These finite sequences are also called **strings.** This string is also denoted by $a_1a_2\ldots a_n$. (Recall that bit strings, which are finite sequences of bits, were introduced in Section 1.1.) The **length** of the string S is the number of terms in this string. The **empty string,** denoted by λ, is the string that has no terms. The empty string has length zero.

EXAMPLE 4 The string $abcd$ is a string of length four. ◀

Special Integer Sequences

A common problem in discrete mathematics is finding a formula or a general rule for constructing the terms of a sequence. Sometimes only a few terms of a sequence solving a problem are known; the goal is to identify the sequence. Even though the initial terms of a sequence do not determine the entire sequence (after all, there are infinitely many different sequences that start with any finite set of initial terms), knowing the first few terms may help you make an educated conjecture about the identity of your sequence. Once you have made this conjecture, you can try to verify that you have the correct sequence.

When trying to deduce a possible formula or rule for the terms of a sequence from the initial terms, try to find a pattern in these terms. You might also see whether you can determine how a term might have been produced from those preceding it. There are many questions you could ask, but some of the more useful are:

- Are there runs of the same value? That is, does the same value occur many times in a row?
- Are terms obtained from previous terms by adding the same amount or an amount that depends on the position in the sequence?
- Are terms obtained from previous terms by multiplying by a particular amount?

- Are terms obtained by combining previous terms in a certain way?
- Are there cycles among the terms?

EXAMPLE 5 Find formulae for the sequences with the following first five terms: (a) 1, 1/2, 1/4, 1/8, 1/16 (b) 1, 3, 5, 7, 9 (c) 1, −1, 1, −1, 1.

Extra Examples

Solution: (a) We recognize that the denominators are powers of 2. The sequence with $a_n = 1/2^n$, $n = 0, 1, 2, \ldots$ is a possible match. This proposed sequence is a geometric progression with $a = 1$ and $r = 1/2$.

(b) We note that each term is obtained by adding 2 to the previous term. The sequence with $a_n = 2n + 1$, $n = 0, 1, 2, \ldots$ is a possible match. This proposed sequence is an arithmetic progression with $a = 1$ and $d = 2$.

(c) The terms alternate between 1 and −1. The sequence with $a_n = (-1)^n$, $n = 0, 1, 2 \ldots$ is a possible match. This proposed sequence is a geometric progression with $a = 1$ and $r = -1$. ◀

Examples 6 and 7 illustrate how we can analyze sequences to find how the terms are constructed.

EXAMPLE 6 How can we produce the terms of a sequence if the first 10 terms are 1, 2, 2, 3, 3, 3, 4, 4, 4, 4?

Solution: Note that the integer 1 appears once, the integer 2 appears twice, the integer 3 appears three times, and the integer 4 appears four times. A reasonable rule for generating this sequence is that the integer n appears exactly n times, so the next five terms of the sequence would all be 5, the following six terms would all be 6, and so on. The sequence generated this way is a possible match. ◀

EXAMPLE 7 How can we produce the terms of a sequence if the first 10 terms are 5, 11, 17, 23, 29, 35, 41, 47, 53, 59?

Solution: Note that each of the first 10 terms of this sequence after the first is obtained by adding 6 to the previous term. (We could see this by noticing that the difference between consecutive terms is 6.) Consequently, the nth term could be produced by starting with 5 and adding 6 a total of $n - 1$ times; that is, a reasonable guess is that the nth term is $5 + 6(n - 1) = 6n - 1$. (This is an arithmetic progression with $a = 5$ and $d = 6$.) ◀

Another useful technique for finding a rule for generating the terms of a sequence is to compare the terms of a sequence of interest with the terms of a well-known integer sequence, such as terms of an arithmetic progression, terms of a geometric progression, perfect squares, perfect cubes, and so on. The first 10 terms of some sequences you may want to keep in mind are displayed in Table 1.

EXAMPLE 8 Conjecture a simple formula for a_n if the first 10 terms of the sequence $\{a_n\}$ are 1, 7, 25, 79, 241, 727, 2185, 6559, 19681, 59047.

Solution: To attack this problem, we begin by looking at the difference of consecutive terms, but we do not see a pattern. When we form the ratio of consecutive terms to see whether each term is a multiple of the previous term, we find that this ratio, although not a constant, is close to 3. So it is reasonable to suspect that the terms of this sequence are generated by a formula involving 3^n. Comparing these terms with the corresponding terms of the sequence $\{3^n\}$, we notice that the nth term is 2 less than the corresponding power of 3. We see that $a_n = 3^n - 2$ for $1 \leq n \leq 10$ and conjecture that this formula holds for all n. ◀

TABLE 1 Some Useful Sequences.

nth Term	*First 10 Terms*
n^2	1, 4, 9, 16, 25, 36, 49, 64, 81, 100, ...
n^3	1, 8, 27, 64, 125, 216, 343, 512, 729, 1000, ...
n^4	1, 16, 81, 256, 625, 1296, 2401, 4096, 6561, 10000, ...
2^n	2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, ...
3^n	3, 9, 27, 81, 243, 729, 2187, 6561, 19683, 59049, ...
$n!$	1, 2, 6, 24, 120, 720, 5040, 40320, 362880, 3628800, ...

We will see throughout this text that integer sequences appear in a wide range of contexts in discrete mathematics. Sequences we have or will encounter include the sequence of prime numbers (Chapter 3), the number of ways to order n discrete objects (Chapter 5), the number of moves required to solve the famous Tower of Hanoi puzzle with n disks (Chapter 7), and the number of rabbits on an island after n months (Chapter 7).

Integer sequences appear in an amazingly wide range of subject areas besides discrete mathematics, including biology, engineering, chemistry, and physics, as well as in puzzles. An amazing database of over 100,000 different integer sequences can be found in the *On-Line Encyclopedia of Integer Sequences*. This database was originated by Neil Sloane in the 1960s. The last printed version of this database was published in 1995 ([SIPI95]); the current encyclopedia would occupy more than 150 volumes of the size of the 1995 book. New sequences are added regularly to this database. There is also a program accessible via the Web that you can use to find sequences from the encyclopedia that match initial terms you provide.

Links

Summations

Next, we introduce **summation notation.** We begin by describing the notation used to express the sum of the terms

$$a_m, a_{m+1}, \ldots, a_n$$

from the sequence $\{a_n\}$. We use the notation

$$\sum_{j=m}^{n} a_j, \qquad \sum_{j=m}^{n} a_j, \qquad \text{or} \qquad \sum_{1 \le j \le n} a_j$$

to represent

$$a_m + a_{m+1} + \cdots + a_n.$$

Here, the variable j is called the **index of summation,** and the choice of the letter j as the variable is arbitrary; that is, we could have used any other letter, such as i or k. Or, in notation,

$$\sum_{j=m}^{n} a_j = \sum_{i=m}^{n} a_i = \sum_{k=m}^{n} a_k.$$

Here, the index of summation runs through all integers starting with its **lower limit** m and ending with its **upper limit** n. A large uppercase Greek letter sigma, $\sum$, is used to denote summation.

The usual laws for arithmetic apply to summations. For example, when a and b are real numbers, we have $\sum_{j=1}^{n}(ax_j + by_j) = a\sum_{y=1}^{n} x_j + b\sum_{j=1}^{n} y_j$, where $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are real numbers. (We do not present a formal proof of this identity here. Such a proof can be constructed using mathematical induction, a proof method we introduce in Chapter 4. The proof also uses the commutative and associative laws for addition and the distributive law of multiplication over addition.)

We give some examples of summation notation.

EXAMPLE 9 Express the sum of the first 100 terms of the sequence $\{a_n\}$, where $a_n = 1/n$ for $n = 1, 2, 3, \ldots$.

Extra Examples

Solution: The lower limit for the index of summation is 1, and the upper limit is 100. We write this sum as

$$\sum_{j=1}^{100} \frac{1}{j}.$$

◀

EXAMPLE 10 What is the value of $\sum_{j=1}^{5} j^2$?

Solution: We have

$$\begin{aligned}\sum_{j=1}^{5} j^2 &= 1^2 + 2^2 + 3^2 + 4^2 + 5^2\\ &= 1 + 4 + 9 + 16 + 25\\ &= 55.\end{aligned}$$

◀

EXAMPLE 11 What is the value of $\sum_{k=4}^{8}(-1)^k$?

Solution: We have

$$\begin{aligned}\sum_{k=4}^{8}(-1)^k &= (-1)^4 + (-1)^5 + (-1)^6 + (-1)^7 + (-1)^8\\ &= 1 + (-1) + 1 + (-1) + 1\\ &= 1.\end{aligned}$$

◀

Links

NEIL SLOANE (BORN 1939) Neil Sloane studied mathematics and electrical engineering at the University of Melbourne on a scholarship from the Australian state telephone company. He mastered many telephone-related jobs, such as erecting telephone poles, in his summer work. After graduating, he designed minimal-cost telephone networks in Australia. In 1962 he came to the United States and studied electrical engineering at Cornell University. His Ph.D. thesis was on what are now called neural networks. He took a job at Bell Labs in 1969, working in many areas, including network design, coding theory, and sphere packing. He now works for AT&T Labs, moving there from Bell Labs when AT&T split up in 1996. One of his favorite problems is the **kissing problem** (a name he coined), which asks how many spheres can be arranged in n dimensions so that they all touch a central sphere of the same size. (In two dimensions the answer is 6, because 6 pennies can be placed so that they touch a central penny. In three dimensions, 12 billiard balls can be placed so that they touch a central billiard ball. Two billiard balls that just touch are said to "kiss," giving rise to the terminology "kissing problem" and "kissing number.") Sloane, together with Andrew Odlyzko, showed that in 8 and 24 dimensions, the optimal kissing numbers are, respectively, 240 and 196,560. The kissing number is known in dimensions 1, 2, 3, 8, and 24, but not in any other dimensions. Sloane's books include *Sphere Packings, Lattices and Groups,* 3d ed., with John Conway; *The Theory of Error-Correcting Codes* with Jessie MacWilliams; *The Encyclopedia of Integer Sequences* with Simon Plouffe; and *The Rock-Climbing Guide to New Jersey Crags* with Paul Nick. The last book demonstrates his interest in rock climbing; it includes more than 50 climbing sites in New Jersey.

Sometimes it is useful to shift the index of summation in a sum. This is often done when two sums need to be added but their indices of summation do not match. When shifting an index of summation, it is important to make the appropriate changes in the corresponding summand. This is illustrated by Example 12.

EXAMPLE 12 Suppose we have the sum

$$\sum_{j=1}^{5} j^2$$

Extra Examples

but want the index of summation to run between 0 and 4 rather than from 1 to 5. To do this, we let $k = j - 1$. Then the new summation index runs from 0 to 4, and the term j^2 becomes $(k+1)^2$. Hence,

$$\sum_{j=1}^{5} j^2 = \sum_{k=0}^{4} (k+1)^2.$$

It is easily checked that both sums are $1 + 4 + 9 + 16 + 25 = 55$. ◀

Sums of terms of geometric progressions commonly arise (such sums are called **geometric series**). Theorem 1 gives us a formula for the sum of terms of a geometric progression.

THEOREM 1 If a and r are real numbers and $r \neq 0$, then

$$\sum_{j=0}^{n} ar^j = \begin{cases} \dfrac{ar^{n+1} - a}{r - 1} & \text{if } r \neq 1 \\ (n+1)a & \text{if } r = 1. \end{cases}$$

Proof: Let

$$S = \sum_{j=0}^{n} ar^j.$$

To compute S, first multiply both sides of the equality by r and then manipulate the resulting sum as follows:

$$\begin{aligned}
rS &= r\sum_{j=0}^{n} ar^j && \text{substituting summation formula for } S \\
&= \sum_{j=0}^{n} ar^{j+1} && \text{by the distributive property} \\
&= \sum_{k=1}^{n+1} ar^k && \text{shifting the index of summation, with } k = j+1 \\
&= \left(\sum_{k=0}^{n} ar^k\right) + (ar^{n+1} - a) && \text{removing } k = n+1 \text{ term and adding } k = 0 \text{ term} \\
&= S + (ar^{n+1} - a) && \text{substituting } S \text{ for summation formula}
\end{aligned}$$

From these equalities, we see that

$$rS = S + (ar^{n+1} - a).$$

Solving for S shows that if $r \neq 1$, then

$$S = \frac{ar^{n+1} - a}{r - 1}.$$

If $r = 1$, then clearly the sum equals $(n + 1)a$. ◁

EXAMPLE 13 Double summations arise in many contexts (as in the analysis of nested loops in computer programs). An example of a double summation is

$$\sum_{i=1}^{4} \sum_{j=1}^{3} ij.$$

To evaluate the double sum, first expand the inner summation and then continue by computing the outer summation:

$$\begin{aligned}\sum_{i=1}^{4} \sum_{j=1}^{3} ij &= \sum_{i=1}^{4} (i + 2i + 3i) \\ &= \sum_{i=1}^{4} 6i \\ &= 6 + 12 + 18 + 24 = 60.\end{aligned}$$

◀

We can also use summation notation to add all values of a function, or terms of an indexed set, where the index of summation runs over all values in a set. That is, we write

$$\sum_{s \in S} f(s)$$

to represent the sum of the values $f(s)$, for all members s of S.

EXAMPLE 14 What is the value of $\sum_{s \in \{0,2,4\}} s$?

Solution: Because $\sum_{s \in \{0,2,4\}} s$ represents the sum of the values of s for all the members of the set $\{0, 2, 4\}$, it follows that

$$\sum_{s \in \{0,2,4\}} s = 0 + 2 + 4 = 6.$$

◀

Certain sums arise repeatedly throughout discrete mathematics. Having a collection of formulae for such sums can be useful; Table 2 provides a small table of formulae for commonly occurring sums.

We derived the first formula in this table in Theorem 1. The next three formulae give us the sum of the first n positive integers, the sum of their squares, and the sum of their cubes. These three formulae can be derived in many different ways (for example, see Exercises 21 and 22 at the end of this section). Also note that each of these formulae, once known, can easily be proved using mathematical induction, the subject of Section 4.1. The last two formulae in the table involve infinite series and will be discussed shortly.

Example 15 illustrates how the formulae in Table 2 can be useful.

TABLE 2 Some Useful Summation Formulae.

Sum	*Closed Form*
$\sum_{k=0}^{n} ar^k \ (r \neq 0)$	$\frac{ar^{n+1}-a}{r-1}, r \neq 1$
$\sum_{k=1}^{n} k$	$\frac{n(n+1)}{2}$
$\sum_{k=1}^{n} k^2$	$\frac{n(n+1)(2n+1)}{6}$
$\sum_{k=1}^{n} k^3$	$\frac{n^2(n+1)^2}{4}$
$\sum_{k=0}^{\infty} x^k, \lvert x\rvert < 1$	$\frac{1}{1-x}$
$\sum_{k=1}^{\infty}, kx^{k-1}, \lvert x\rvert < 1$	$\frac{1}{(1-x)^2}$

EXAMPLE 15 Find $\sum_{k=50}^{100} k^2$.

Solution: First note that because $\sum_{k=1}^{100} k^2 = \sum_{k=1}^{49} k^2 + \sum_{k=50}^{100} k^2$, we have

$$\sum_{k=50}^{100} k^2 = \sum_{k=1}^{100} k^2 - \sum_{k=1}^{49} k^2.$$

Using the formula $\sum_{k=1}^{n} k^2 = n(n+1)(2n+1)/6$ from Table 2, we see that

$$\sum_{k=50}^{100} k^2 = \frac{100 \cdot 101 \cdot 201}{6} - \frac{49 \cdot 50 \cdot 99}{6} = 338{,}350 - 40{,}425 = 297{,}925.$$

◀

SOME INFINITE SERIES Although most of the summations in this book are finite sums, infinite series are important in some parts of discrete mathematics. Infinite series are usually studied in a course in calculus and even the definition of these series requires the use of calculus, but sometimes they arise in discrete mathematics, because discrete mathematics deals with infinite collections of discrete elements. In particular, in our future studies in discrete mathematics, we will find the closed forms for the infinite series in Examples 16 and 17 to be quite useful.

EXAMPLE 16 (*Requires calculus*) Let x be a real number with $|x| < 1$. Find $\sum_{n=0}^{\infty} x^n$.

Extra Examples

Solution: By Theorem 1 with $a = 1$ and $r = x$ we see that $\sum_{n=0}^{k} x^n = \frac{x^{k+1}-1}{x-1}$. Because $|x| < 1$, x^{k+1} approaches 0 as k approaches infinity. It follows that

$$\sum_{n=0}^{\infty} x^n = \lim_{k\to\infty} \frac{x^{k+1}-1}{x-1} = \frac{-1}{x-1} = \frac{1}{1-x}.$$

◀

We can produce new summation formulae by differentiating or integrating existing formulae.

EXAMPLE 17 (*Requires calculus*) Differentiating both sides of the equation

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x},$$

from Example 16 we find that

$$\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}.$$

(This differentiation is valid for $|x| < 1$ by a theorem about infinite series.) ◀

Cardinality

In Section 2.1 we defined the cardinality of a finite set to be the number of elements in the set. That is, the cardinality of a finite set tells us when two finite sets are the same size, or when one is bigger than the other. Can we extend this notion to an infinite set? Recall from Exercise 75 in Section 2.3 that there is a one-to-one correspondence between any two finite sets with the same number of elements. This observation lets us extend the concept of cardinality to all sets, both finite and infinite, with Definition 4.

DEFINITION 4 The sets A and B have the same *cardinality* if and only if there is a one-to-one correspondence from A to B.

We will now split infinite sets into two groups, those with the same cardinality as the set of natural numbers and those with different cardinality.

DEFINITION 5 A set that is either finite or has the same cardinality as the set of positive integers is called *countable*. A set that is not countable is called *uncountable*. When an infinite set S is countable, we denote the cardinality of S by $\aleph_0$ (where $\aleph$ is aleph, the first letter of the Hebrew alphabet). We write $|S| = \aleph_0$ and say that S has cardinality "aleph null."

We now give examples of countable and uncountable sets.

EXAMPLE 18 Show that the set of odd positive integers is a countable set.

Solution: To show that the set of odd positive integers is countable, we will exhibit a one-to-one correspondence between this set and the set of positive integers. Consider the function

$$f(n) = 2n - 1$$

from $\mathbf{Z}^+$ to the set of odd positive integers. We show that f is a one-to-one correspondence by showing that it is both one-to-one and onto. To see that it is one-to-one, suppose that $f(n) = f(m)$. Then $2n - 1 = 2m - 1$, so $n = m$. To see that it is onto, suppose that t is an odd positive

FIGURE 1 A One-to-One Correspondence Between $\mathbf{Z}^+$ and the Set of Odd Positive Integers.

FIGURE 2 The Positive Rational Numbers Are Countable.

integer. Then t is 1 less than an even integer $2k$, where k is a natural number. Hence $t = 2k - 1 = f(k)$. We display this one-to-one correspondence in Figure 1. ◀

An infinite set is countable if and only if it is possible to list the elements of the set in a sequence (indexed by the positive integers). The reason for this is that a one-to-one correspondence f from the set of positive integers to a set S can be expressed in terms of a sequence $a_1, a_2, \ldots, a_n, \ldots$, where $a_1 = f(1)$, $a_2 = f(2), \ldots, a_n = f(n), \ldots$. For instance, the set of odd integers can be listed in a sequence $a_1, a_2, \ldots, a_n, \ldots$, where $a_n = 2n - 1$.

We can show that the set of all integers is countable by listing its members.

EXAMPLE 19 Show that the set of all integers is countable.

Solution: We can list all integers in a sequence by starting with 0 and alternating between positive and negative integers: $0, 1, -1, 2, -2, \ldots$. Alternately, we could find a one-to-one correspondence between the set of positive integers and the set of all integers. We leave it to the reader to show that the function $f(n) = n/2$ when n is even and $f(n) = -(n-1)/2$ when n is odd is such a function. Consequently, the set of all integers is countable. ◀

It is not surprising that the set of odd integers and the set of all integers are both countable sets. Many people are amazed to learn that the set of rational numbers is countable, as Example 20 demonstrates.

EXAMPLE 20 Show that the set of positive rational numbers is countable.

Solution: It may seem surprising that the set of positive rational numbers is countable, but we will show how we can list the positive rational numbers as a sequence $r_1, r_2, \ldots, r_n, \ldots$. First, note that every positive rational number is the quotient p/q of two positive integers. We can arrange the positive rational numbers by listing those with denominator $q = 1$ in the first row, those with denominator $q = 2$ in the second row, and so on, as displayed in Figure 2.

The key to listing the rational numbers in a sequence is to first list the positive rational numbers p/q with $p + q = 2$, followed by those with $p + q = 3$, followed by those with $p + q = 4$, and so on, following the path shown in Figure 2. Whenever we encounter a number p/q that is already listed, we do not list it again. For example, when we come to $2/2 = 1$ we do not list it because we have already listed $1/1 = 1$. The initial terms in the list of positive rational numbers we have constructed are $1, 1/2, 2, 3, 1/3, 1/4, 2/3, 3/2, 4, 5$, and so on. These numbers are shown circled; the uncircled numbers in the list are those we leave out

because they are already listed. Because all positive rational numbers are listed once, as the reader can verify, we have shown that the set of positive rational numbers is countable. ◀

We have seen that the set of rational numbers is a countable set. Do we have a promising candidate for an uncountable set? The first place we might look is the set of real numbers. In Example 21 we use an important proof method, introduced in 1879 by Georg Cantor and known as the **Cantor diagonalization argument,** to prove that the set of real numbers is not countable. This proof method is used extensively in mathematical logic and in the theory of computation.

Links

EXAMPLE 21 Show that the set of real numbers is an uncountable set.

Extra Examples

Solution: To show that the set of real numbers is uncountable, we suppose that the set of real numbers is countable and arrive at a contradiction. Then, the subset of all real numbers that fall between 0 and 1 would also be countable (because any subset of a countable set is also countable; see Exercise 36 at the end of the section). Under this assumption, the real numbers between 0 and 1 can be listed in some order, say, $r_1, r_2, r_3, \ldots$. Let the decimal representation of these real numbers be

$$\begin{aligned} r_1 &= 0.d_{11}d_{12}d_{13}d_{14}\ldots \\ r_2 &= 0.d_{21}d_{22}d_{23}d_{24}\ldots \\ r_3 &= 0.d_{31}d_{32}d_{33}d_{34}\ldots \\ r_4 &= 0.d_{41}d_{42}d_{43}d_{44}\ldots \\ &\vdots \end{aligned}$$

where $d_{ij} \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. (For example, if $r_1 = 0.23794102\ldots$, we have $d_{11} = 2$, $d_{12} = 3$, $d_{13} = 7$, and so on.) Then, form a new real number with decimal expansion $r = 0.d_1d_2d_3d_4\ldots$, where the decimal digits are determined by the following rule:

$$d_i = \begin{cases} 4 & \text{if } d_{ii} \neq 4 \\ 5 & \text{if } d_{ii} = 4. \end{cases}$$

(As an example, suppose that $r_1 = 0.23794102\ldots$, $r_2 = 0.44590138\ldots$, $r_3 = 0.09118764\ldots$, $r_4 = 0.80553900\ldots$, and so on. Then we have $r = 0.d_1d_2d_3d_4\ldots = 0.4544\ldots$, where $d_1 = 4$ because $d_{11} \neq 4$, $d_2 = 5$ because $d_{22} = 4$, $d_3 = 4$ because $d_{33} \neq 4$, $d_4 = 4$ because $d_{44} \neq 4$, and so on.)

Every real number has a unique decimal expansion (when the possibility that the expansion has a tail end that consists entirely of the digit 9 is excluded). Then, the real number r is not equal to any of $r_1, r_2, \ldots$ because the decimal expansion of r differs from the decimal expansion of r_i in the ith place to the right of the decimal point, for each i.

Because there is a real number r between 0 and 1 that is not in the list, the assumption that all the real numbers between 0 and 1 could be listed must be false. Therefore, all the real numbers between 0 and 1 cannot be listed, so the set of real numbers between 0 and 1 is uncountable. Any set with an uncountable subset is uncountable (see Exercise 37 at the end of this section). Hence, the set of real numbers is uncountable. ◀

Exercises

1. Find these terms of the sequence $\{a_n\}$, where $a_n = 2 \cdot (-3)^n + 5^n$.
a) a_0 **b)** a_1 **c)** a_4 **d)** a_5

2. What is the term a_8 of the sequence $\{a_n\}$ if a_n equals
a) 2^{n-1}? **b)** 7?
c) $1 + (-1)^n$? **d)** $-(-2)^n$?

3. What are the terms a_0, a_1, a_2, and a_3 of the sequence $\{a_n\}$, where a_n equals
   a) $2^n + 1$?
   b) $(n+1)^{n+1}$?
   c) $\lfloor n/2 \rfloor$?
   d) $\lfloor n/2 \rfloor + \lceil n/2 \rceil$?
4. What are the terms a_0, a_1, a_2, and a_3 of the sequence $\{a_n\}$, where a_n equals
   a) $(-2)^n$?
   b) 3?
   c) $7 + 4^n$?
   d) $2^n + (-2)^n$?
5. List the first 10 terms of each of these sequences.
   a) the sequence that begins with 2 and in which each successive term is 3 more than the preceding term
   b) the sequence that lists each positive integer three times, in increasing order
   c) the sequence that lists the odd positive integers in increasing order, listing each odd integer twice
   d) the sequence whose nth term is $n! - 2^n$
   e) the sequence that begins with 3, where each succeeding term is twice the preceding term
   f) the sequence whose first two terms are 1 and each succeeding term is the sum of the two preceding terms (This is the famous Fibonacci sequence, which we will study later in this text.)
   g) the sequence whose nth term is the number of bits in the binary expansion of the number n (defined in Section 3.6)
   h) the sequence where the nth term is the number of letters in the English word for the index n
6. List the first 10 terms of each of these sequences.
   a) the sequence obtained by starting with 10 and obtaining each term by subtracting 3 from the previous term
   b) the sequence whose nth term is the sum of the first n positive integers
   c) the sequence whose nth term is $3^n - 2^n$
   d) the sequence whose nth term is $\lfloor \sqrt{n} \rfloor$
   e) the sequence whose first two terms are 1 and 2 and each succeeding term is the sum of the two previous terms
   f) the sequence whose nth term is the largest integer whose binary expansion (defined in Section 3.6) has n bits (Write your answer in decimal notation.)
   g) the sequence whose terms are constructed sequentially as follows: start with 1, then add 1, then multiply by 1, then add 2, then multiply by 2, and so on
   h) the sequence whose nth term is the largest integer k such that $k! \leq n$
7. Find at least three different sequences beginning with the terms 1, 2, 4 whose terms are generated by a simple formula or rule.
8. Find at least three different sequences beginning with the terms 3, 5, 7 whose terms are generated by a simple formula or rule.
9. For each of these lists of integers, provide a simple formula or rule that generates the terms of an integer sequence that begins with the given list. Assuming that your formula or rule is correct, determine the next three terms of the sequence.
   a) 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, ...
   b) 1, 2, 2, 3, 4, 4, 5, 6, 6, 7, 8, 8, ...
   c) 1, 0, 2, 0, 4, 0, 8, 0, 16, 0, ...
   d) 3, 6, 12, 24, 48, 96, 192, ...
   e) 15, 8, 1, −6, −13, −20, −27, ...
   f) 3, 5, 8, 12, 17, 23, 30, 38, 47, ...
   g) 2, 16, 54, 128, 250, 432, 686, ...
   h) 2, 3, 7, 25, 121, 721, 5041, 40321, ...
10. For each of these lists of integers, provide a simple formula or rule that generates the terms of an integer sequence that begins with the given list. Assuming that your formula or rule is correct, determine the next three terms of the sequence.
   a) 3, 6, 11, 18, 27, 38, 51, 66, 83, 102, ...
   b) 7, 11, 15, 19, 23, 27, 31, 35, 39, 43, ...
   c) 1, 10, 11, 100, 101, 110, 111, 1000, 1001, 1010, 1011, ...
   d) 1, 2, 2, 2, 3, 3, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, ...
   e) 0, 2, 8, 26, 80, 242, 728, 2186, 6560, 19682, ...
   f) 1, 3, 15, 105, 945, 10395, 135135, 2027025, 34459425, ...
   g) 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, ...
   h) 2, 4, 16, 256, 65536, 4294967296, ...

**11. Show that if a_n denotes the nth positive integer that is not a perfect square, then $a_n = n + \{\sqrt{n}\}$, where $\{x\}$ denotes the integer closest to the real number x.

*12. Let a_n be the nth term of the sequence 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, ..., constructed by including the integer k exactly k times. Show that $a_n = \lfloor \sqrt{2n} + \frac{1}{2} \rfloor$.

13. What are the values of these sums?
   a) $\sum_{k=1}^{5} (k+1)$
   b) $\sum_{j=0}^{4} (-2)^j$
   c) $\sum_{i=1}^{10} 3$
   d) $\sum_{j=0}^{8} (2^{j+1} - 2^j)$
14. What are the values of these sums, where $S = \{1, 3, 5, 7\}$?
   a) $\sum_{j \in S} j$
   b) $\sum_{j \in S} j^2$
   c) $\sum_{j \in S} (1/j)$
   d) $\sum_{j \in S} 1$
15. What is the value of each of these sums of terms of a geometric progression?
   a) $\sum_{j=0}^{8} 3 \cdot 2^j$
   b) $\sum_{j=1}^{8} 2^j$
   c) $\sum_{j=2}^{8} (-3)^j$
   d) $\sum_{j=0}^{8} 2 \cdot (-3)^j$
16. Find the value of each of these sums.
   a) $\sum_{j=0}^{8} (1 + (-1)^j)$
   b) $\sum_{j=0}^{8} (3^j - 2^j)$

c) $\sum_{j=0}^{8}(2 \cdot 3^j + 3 \cdot 2^j)$ d) $\sum_{j=0}^{8}(2^{j+1} - 2^j)$

17. Compute each of these double sums.

a) $\sum_{i=1}^{2}\sum_{j=1}^{3}(i+j)$ b) $\sum_{i=0}^{2}\sum_{j=0}^{3}(2i+3j)$

c) $\sum_{i=1}^{3}\sum_{j=0}^{2} i$ d) $\sum_{i=0}^{2}\sum_{j=1}^{3} ij$

18. Compute each of these double sums.

a) $\sum_{i=1}^{3}\sum_{j=1}^{2}(i-j)$ b) $\sum_{i=0}^{3}\sum_{j=0}^{2}(3i+2j)$

c) $\sum_{i=1}^{3}\sum_{j=0}^{2} j$ d) $\sum_{i=0}^{2}\sum_{j=0}^{3} i^2 j^3$

19. Show that $\sum_{j=1}^{n}(a_j - a_{j-1}) = a_n - a_0$, where $a_0, a_1, \ldots, a_n$ is a sequence of real numbers. This type of sum is called **telescoping.**

20. Use the identity $1/(k(k+1)) = 1/k - 1/(k+1)$ and Exercise 19 to compute $\sum_{k=1}^{n} 1/(k(k+1))$.

21. Sum both sides of the identity $k^2 - (k-1)^2 = 2k - 1$ from $k = 1$ to $k = n$ and use Exercise 19 to find

a) a formula for $\sum_{k=1}^{n}(2k-1)$ (the sum of the first n odd natural numbers).

b) a formula for $\sum_{k=1}^{n} k$.

***22.** Use the technique given in Exercise 19, together with the result of Exercise 21b, to derive the formula for $\sum_{k=1}^{n} k^2$ given in Table 2. (*Hint*: Take $a_k = k^3$ in the telescoping sum in Exercise 19.)

23. Find $\sum_{k=100}^{200} k$. (Use Table 2.)

24. Find $\sum_{k=99}^{200} k^3$. (Use Table 2.)

***25.** Find a formula for $\sum_{k=0}^{m} \lfloor \sqrt{k} \rfloor$, when m is a positive integer.

***26.** Find a formula for $\sum_{k=0}^{m} \lfloor \sqrt[3]{k} \rfloor$, when m is a positive integer.

There is also a special notation for products. The product of $a_m, a_{m+1}, \ldots, a_n$ is represented by

$$\prod_{j=m}^{n} a_j.$$

27. What are the values of the following products?

a) $\prod_{i=0}^{10} i$ b) $\prod_{i=5}^{8} i$

c) $\prod_{i=1}^{100}(-1)^i$ d) $\prod_{i=1}^{10} 2$

Recall that the value of the factorial function at a positive integer n, denoted by $n!$, is the product of the positive integers from 1 to n, inclusive. Also, we specify that $0! = 1$.

28. Express $n!$ using product notation.

29. Find $\sum_{j=0}^{4} j!$.

30. Find $\prod_{j=0}^{4} j!$.

31. Determine whether each of these sets is countable or uncountable. For those that are countable, exhibit a one-to-one correspondence between the set of natural numbers and that set.

a) the negative integers

b) the even integers

c) the real numbers between 0 and $\frac{1}{2}$

d) integers that are multiples of 7

32. Determine whether each of these sets is countable or uncountable. For those that are countable, exhibit a one-to-one correspondence between the set of natural numbers and that set.

a) the integers greater than 10

b) the odd negative integers

c) the real numbers between 0 and 2

d) integers that are multiples of 10

33. Determine whether each of these sets is countable or uncountable. For those that are countable, exhibit a one-to-one correspondence between the set of natural numbers and that set.

a) all bit strings not containing the bit 0

b) all positive rational numbers that cannot be written with denominators less than 4

c) the real numbers not containing 0 in their decimal representation

d) the real numbers containing only a finite number of 1s in their decimal representation

34. Determine whether each of these sets is countable or uncountable. For those that are countable, exhibit a one-to-one correspondence between the set of natural numbers and that set.

a) integers not divisible by 3

b) integers divisible by 5 but not by 7

c) the real numbers with decimal representations consisting of all 1s

d) the real numbers with decimal representations of all 1s or 9s

35. If A is an uncountable set and B is a countable set, must $A - B$ be uncountable?

36. Show that a subset of a countable set is also countable.

37. Show that if A and B are sets, A is uncountable, and $A \subseteq B$, then B is uncountable.

38. Show that if A and B are sets with the same cardinality, then the power set of A and the power set of B have the same cardinality.

39. Show that if A and B are sets with the same cardinality and C and D are sets with the same cardinality, then $A \times C$ and $B \times D$ have the same cardinality.

40. Show that the union of two countable sets is countable.

****41.** Show that the union of a countable number of countable sets is countable.

42. Show that the set $\mathbf{Z}^+ \times \mathbf{Z}^+$ is countable.

***43.** Show that the set of all finite bit strings is countable.

***44.** Show that the set of real numbers that are solutions of quadratic equations $ax^2 + bx + c = 0$, where a, b, and c are integers, is countable.

***45.** Show that the set of all computer programs in a particular programming language is countable. [*Hint:* A computer program written in a programming language can be thought of as a string of symbols from a finite alphabet.]

***46.** Show that the set of functions from the positive integers to the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ is uncountable. [*Hint:* First set up a one-to-one correspondence between the set of real numbers between 0 and 1 and a subset of these functions. Do this by associating to the real number $0.d_1d_2\ldots d_n\ldots$ the function f with $f(n) = d_n$.]

***47.** We say that a function is **computable** if there is a computer program that finds the values of this function. Use Exercises 45 and 46 to show that there are functions that are not computable.

***48.** Show that if S is a set, then there does not exist an onto function f from S to $P(S)$, the power set of S. [*Hint:* Suppose such a function f existed. Let $T = \{s \in S \mid s \notin f(s)\}$ and show that no element s can exist for which $f(s) = T$.]

Key Terms and Results

TERMS

set: a collection of distinct objects
axiom: a basic assumption of a theory
paradox: a logical inconsistency
element, member of a set: an object in a set
$\emptyset$ (empty set, null set): the set with no members
universal set: the set containing all objects under consideration
Venn diagram: a graphical representation of a set or sets
$S = T$ (set equality): S and T have the same elements
$S \subseteq T$ (S is a subset of T): every element of S is also an element of T
$S \subset T$ (S is a proper subset of T): S is a subset of T and $S \neq T$
finite set: a set with n elements, where n is a nonnegative integer
infinite set: a set that is not finite
$|S|$ (the cardinality of S): the number of elements in S
$P(S)$ (the power set of S): the set of all subsets of S
$A \cup B$ (the union of A and B): the set containing those elements that are in at least one of A and B
$A \cap B$ (the intersection of A and B): the set containing those elements that are in both A and B.
$A - B$ (the difference of A and B): the set containing those elements that are in A but not in B
$\overline{A}$ (the complement of A): the set of elements in the universal set that are not in A
$A \oplus B$ (the symmetric difference of A and B): the set containing those elements in exactly one of A and B
membership table: a table displaying the membership of elements in sets
function from A to B: an assignment of exactly one element of B to each element of A
domain of f: the set A, where f is a function from A to B
codomain of f: the set B, where f is a function from A to B
b is the image of a under f: $b = f(a)$
a is a pre-image of b under f: $f(a) = b$
range of f: the set of images of f
onto function, surjection: a function from A to B such that every element of B is the image of some element in A
one-to-one function, injection: a function such that the images of elements in its domain are all different
one-to-one correspondence, bijection: a function that is both one-to-one and onto
inverse of f: the function that reverses the correspondence given by f (when f is a bijection)
$f \circ g$ (composition of f and g): the function that assigns $f(g(x))$ to x
$\lfloor x \rfloor$ (floor function): the largest integer not exceeding x
$\lceil x \rceil$ (ceiling function): the smallest integer greater than or equal to x
sequence: a function with domain that is a subset of the set of integers
geometric progression: a sequence of the form $a, ar, ar^2, \ldots,$ where a and r are real numbers
arithmetic progression: a sequence of the form a, $a + d$, $a + 2d, \ldots,$ where a and d are real numbers
string: a finite sequence
empty string: a string of length zero
$\sum_{i=1}^{n} a_i$: the sum $a_1 + a_2 + \cdots + a_n$
$\prod_{i=1}^{n} a_i$: the product $a_1a_2 \cdots a_n$
countable set: a set that either is finite or can be placed in one-to-one correspondence with the set of positive integers
uncountable set: a set that is not countable
Cantor diagonalization argument: a proof technique that can be used to show that the set of real numbers is uncountable

RESULTS

The set identities given in Table 1 in Section 2.2
The summation formulae in Table 2 in Section 2.4
The set of rational numbers is countable.
The set of real numbers is uncountable.

Review Questions

1. Explain what it means for one set to be a subset of another set. How do you prove that one set is a subset of another set?
2. What is the empty set? Show that the empty set is a subset of every set.
3. a) Define $|S|$, the cardinality of the set S.
   b) Give a formula for $|A \cup B|$, where A and B are sets.
4. a) Define the power set of a set S.
   b) When is the empty set in the power set of a set S?
   c) How many elements does the power set of a set S with n elements have?
5. a) Define the union, intersection, difference, and symmetric difference of two sets.
   b) What are the union, intersection, difference, and symmetric difference of the set of positive integers and the set of odd integers?
6. a) Explain what it means for two sets to be equal.
   b) Describe as many of the ways as you can to show that two sets are equal.
   c) Show in at least two different ways that the sets $A - (B \cap C)$ and $(A - B) \cup (A - C)$ are equal.
7. Explain the relationship between logical equivalences and set identities.
8. a) Define the domain, codomain, and the range of a function.
   b) Let $f(n)$ be the function from the set of integers to the set of integers such that $f(n) = n^2 + 1$. What are the domain, codomain, and range of this function?
9. a) Define what it means for a function from the set of positive integers to the set of positive integers to be one-to-one.
   b) Define what it means for a function from the set of positive integers to the set of positive integers to be onto.
   c) Give an example of a function from the set of positive integers to the set of positive integers that is both one-to-one and onto.
   d) Give an example of a function from the set of positive integers to the set of positive integers that is one-to-one but not onto.
   e) Give an example of a function from the set of positive integers to the set of positive integers that is not one-to-one but is onto.
   f) Give an example of a function from the set of positive integers to the set of positive integers that is neither one-to-one nor onto.
10. a) Define the inverse of a function.
    b) When does a function have an inverse?
    c) Does the function $f(n) = 10 - n$ from the set of integers to the set of integers have an inverse? If so, what is it?
11. a) Define the floor and ceiling functions from the set of real numbers to the set of integers.
    b) For which real numbers x is it true that $\lfloor x \rfloor = \lceil x \rceil$?
12. Conjecture a formula for the terms of the sequence that begins 8, 14, 32, 86, 248 and find the next three terms of your sequence.
13. What is the sum of the terms of the geometric progression $a + ar + \cdots + ar^n$ when $r \neq 1$?
14. Show that the set of odd integers is countable.
15. Give an example of an uncountable set.

Supplementary Exercises

1. Let A be the set of English words that contain the letter x, and let B be the set of English words that contain the letter q. Express each of these sets as a combination of A and B.
   a) The set of English words that do not contain the letter x.
   b) The set of English words that contain both an x and a q.
   c) The set of English words that contain an x but not a q.
   d) The set of English words that do not contain either an x or a q.
   e) The set of English words that contain an x or a q, but not both.
2. Show that if A is a subset of B, then the power set of A is a subset of the power set of B.
3. Suppose that A and B are sets such that the power set of A is a subset of the power set of B. Does it follow that A is a subset of B?
4. Let $\mathbf{E}$ denote the set of even integers and $\mathbf{O}$ denote the set of odd integers. As usual, let $\mathbf{Z}$ denote the set of all integers. Determine each of these sets.
   a) $\mathbf{E} \cup \mathbf{O}$ b) $\mathbf{E} \cap \mathbf{O}$ c) $\mathbf{Z} - \mathbf{E}$ d) $\mathbf{Z} - \mathbf{O}$
5. Show that if A and B are sets, then $A - (A - B) = A \cap B$.
6. Let A and B be sets. Show that $A \subseteq B$ if and only if $A \cap B = A$.
7. Let A, B, and C be sets. Show that $(A - B) - C$ is not necessarily equal to $A - (B - C)$.
8. Suppose that A, B, and C are sets. Prove or disprove that $(A - B) - C = (A - C) - B$.
9. Suppose that A, B, C, and D are sets. Prove or disprove that $(A - B) - (C - D) = (A - C) - (B - D)$.
10. Show that if A and B are finite sets, then $|A \cap B| \leq |A \cup B|$. Determine when this relationship is an equality.

11. Let A and B be sets in a finite universal set U. List the following in order of increasing size.
    a) $|A|, |A \cup B|, |A \cap B|, |U|, |\emptyset|$
    b) $|A - B|, |A \oplus B|, |A| + |B|, |A \cup B|, |\emptyset|$
12. Let A and B be subsets of the finite universal set U. Show that $|\overline{A} \cap \overline{B}| = |U| - |A| - |B| + |A \cap B|$.
13. Let f and g be functions from $\{1, 2, 3, 4\}$ to $\{a, b, c, d\}$ and from $\{a, b, c, d\}$ to $\{1, 2, 3, 4\}$, respectively, such that $f(1) = d$, $f(2) = c$, $f(3) = a$, and $f(4) = b$, and $g(a) = 2$, $g(b) = 1$, $g(c) = 3$, and $g(d) = 2$.
    a) Is f one-to-one? Is g one-to-one?
    b) Is f onto? Is g onto?
    c) Does either f or g have an inverse? If so, find this inverse.
14. Let f be a one-to-one function from the set A to the set B. Let S and T be subsets of A. Show that $f(S \cap T) = f(S) \cap f(T)$.
15. Give an example to show that the equality in Exercise 14 may not hold if f is not one-to-one.

Suppose that f is a function from A to B. We define the function S_f from $P(A)$ to $P(B)$ by the rule $S_f(X) = f(X)$ for each subset X of A. Similarly, we define the function $S_{f^{-1}}$ from $P(B)$ to $P(A)$ by the rule $S_{f^{-1}}(Y) = f^{-1}(Y)$ for each subset Y of B. Here, we are using Definition 4, and the definition of the inverse image of a set found in the preamble to Exercise 38, both in Section 2.3.

*16. a) Prove that if f is a one-to-one function from A to B, then S_f is a one-to-one function from $P(A)$ to $P(B)$.
    b) Prove that if f is an onto function from A to B, then S_f is an onto function from $P(A)$ to $P(B)$.
    c) Prove that if f is an onto function from A to B, then $S_{f^{-1}}$ is a one-to-one function from $P(B)$ to $P(A)$.
    d) Prove that if f is a one-to-one function from A to B, then $S_{f^{-1}}$ is an onto function from $P(B)$ to $P(A)$.
    e) Use parts (a) through (d) to conclude that if f is a one-to-one correspondence from A to B, then S_f is a one-to-one correspondence from $P(A)$ to $P(B)$ and $S_{f^{-1}}$ is a one-to-one correspondence from $P(B)$ to $P(A)$.
17. Prove that if f and g are functions from A to B and $S_f = S_g$ (using the definition in the preamble to Exercise 16), then $f(x) = g(x)$ for all $x \in A$.
18. Show that if n is an integer, then $n = \lceil n/2 \rceil + \lfloor n/2 \rfloor$.
19. For which real numbers x and y is it true that $\lfloor x + y \rfloor = \lfloor x \rfloor + \lfloor y \rfloor$?
20. For which real numbers x and y is it true that $\lceil x + y \rceil = \lceil x \rceil + \lceil y \rceil$?
21. For which real numbers x and y is it true that $\lceil x + y \rceil = \lceil x \rceil + \lfloor y \rfloor$?
22. Prove that $\lfloor n/2 \rfloor \lceil n/2 \rceil = \lfloor n^2/4 \rfloor$ for all integers n.
23. Prove that if m is an integer, then $\lfloor x \rfloor + \lfloor m - x \rfloor = m - 1$, unless x is an integer, in which case, it equals m.
24. Prove that if x is a real number, then $\lfloor \lfloor x/2 \rfloor / 2 \rfloor = \lfloor x/4 \rfloor$.
25. Prove that if n is an odd integer, then $\lceil n^2/4 \rceil = (n^2 + 3)/4$.
26. Prove that if m and n are positive integers and x is a real number, then
$$\left\lfloor \frac{\lfloor x \rfloor + n}{m} \right\rfloor = \left\lfloor \frac{x + n}{m} \right\rfloor.$$
*27. Prove that if m is a positive integer and x is a real number, then
$$\lfloor mx \rfloor = \lfloor x \rfloor + \left\lfloor x + \frac{1}{m} \right\rfloor + \left\lfloor x + \frac{2}{m} \right\rfloor + \cdots + \left\lfloor x + \frac{m-1}{m} \right\rfloor.$$
*28. We define the **Ulam numbers** by setting $u_1 = 1$ and $u_2 = 2$. Furthermore, after determining whether the integers less than n are Ulam numbers, we set n equal to the next Ulam number if it can be written uniquely as the sum of two different Ulam numbers. Note that $u_3 = 3$, $u_4 = 4$, $u_5 = 6$, and $u_6 = 8$.
    a) Find the first 20 Ulam numbers.
    b) Prove that there are infinitely many Ulam numbers.
29. Determine the value of $\prod_{k=1}^{100} \frac{k+1}{k}$. (The notation used here for products is defined in the preamble to Exercise 27 in Section 2.4.)
*30. Determine a rule for generating the terms of the sequence that begins $1, 3, 4, 8, 15, 27, 50, 92, \ldots$, and find the next four terms of the sequence.
*31. Determine a rule for generating the terms of the sequence that begins $2, 3, 3, 5, 10, 13, 39, 43, 172, 177, 885, 891, \ldots$, and find the next four terms of the sequence.
*32. Prove that if A and B are countable sets, then $A \times B$ is also a countable set.

Computer Projects

Write programs with the specified input and output.

1. Given subsets A and B of a set with n elements, use bit strings to find $\overline{A}$, $A \cup B$, $A \cap B$, $A - B$, and $A \oplus B$.
2. Given multisets A and B from the same universal set, find $A \cup B$, $A \cap B$, $A - B$, and $A + B$ (see preamble to Exercise 59 of Section 2.2).

3. Given fuzzy sets A and B, find $\overline{A}$, $A \cup B$, and $A \cap B$ (see preamble to Exercise 61 of Section 2.2).
4. Given a function f from $\{1, 2, \ldots, n\}$ to the set of integers, determine whether f is one-to-one.
5. Given a function f from $\{1, 2, \ldots, n\}$ to itself, determine whether f is onto.
6. Given a bijection f from the set $\{1, 2, \ldots, n\}$ to itself, find f^{-1}.

Computations and Explorations

Use a computational program or programs you have written to do these exercises.

1. Given two finite sets, list all elements in the Cartesian product of these two sets.
2. Given a finite set, list all elements of its power set.
3. Calculate the number of one-to-one functions from a set S to a set T, where S and T are finite sets of various sizes. Can you determine a formula for the number of such functions? (We will find such a formula in Chapter 5.)
4. Calculate the number of onto functions from a set S to a set T, where S and T are finite sets of various sizes. Can you determine a formula for the number of such functions? (We will find such a formula in Chapter 7.)

*5. Develop a collection of different rules for generating the terms of a sequence and a program for randomly selecting one of these rules and the particular sequence generated using these rules. Make this part of an interactive program that prompts for the next term of the sequence and determine whether the response is the intended next term.

Writing Projects

Respond to these with essays using outside sources.

1. Discuss how an axiomatic set theory can be developed to avoid Russell's paradox. (See Exercise 38 of Section 2.1.)
2. Research where the concept of a function first arose, and describe how this concept was first used.
3. Explain the different ways in which the *Encyclopedia of Integer Sequences* has been found useful. Also, describe a few of the more unusual sequences in this encyclopedia and how they arise.
4. Define the recently invented EKG sequence and describe some of its properties and open questions about it.
5. Look up the definition of a transcendental number. Explain how to show that such numbers exist and how such numbers can be constructed. Which famous numbers can be shown to be transcendental and for which famous numbers is it still unknown whether they are transcendental?
6. Discuss infinite cardinal numbers and the continuum hypothesis.

CHAPTER

7 Advanced Counting Techniques

7.1 Recurrence Relations

7.2 Solving Linear Recurrence Relations

7.3 Divide-and-Conquer Algorithms and Recurrence Relations

7.4 Generating Functions

7.5 Inclusion–Exclusion

7.6 Applications of Inclusion–Exclusion

Many counting problems cannot be solved easily using the methods discussed in Chapter 5. One such problem is: How many bit strings of length n do not contain two consecutive zeros? To solve this problem, let a_n be the number of such strings of length n. An argument can be given that shows $a_{n+1} = a_n + a_{n-1}$. This equation, called a recurrence relation, and the initial conditions $a_1 = 2$ and $a_2 = 3$ determine the sequence $\{a_n\}$. Moreover, an explicit formula can be found for a_n from the equation relating the terms of the sequence. As we will see, a similar technique can be used to solve many different types of counting problems.

We will also see that many counting problems can be solved using formal power series, called generating functions, where the coefficients of powers of x represent terms of the sequence we are interested in. Besides solving counting problems, we will also be able to use generating functions to solve recurrence relations and to prove combinatorial identities.

Many other kinds of counting problems cannot be solved using the techniques discussed in Chapter 5, such as: How many ways are there to assign seven jobs to three employees so that each employee is assigned at least one job? How many primes are there less than 1000? Both of these problems can be solved by counting the number of elements in the union of sets. We will develop a technique, called the principle of inclusion–exclusion, that counts the number of elements in a union of sets, and we will show how this principle can be used to solve counting problems.

The techniques studied in this chapter, together with the basic techniques of Chapter 5, can be used to solve many counting problems.

7.1 Recurrence Relations

Introduction

The number of bacteria in a colony doubles every hour. If a colony begins with five bacteria, how many will be present in n hours? To solve this problem, let a_n be the number of bacteria at the end of n hours. Because the number of bacteria doubles every hour, the relationship $a_n = 2a_{n-1}$ holds whenever n is a positive integer. This relationship, together with the initial condition $a_0 = 5$, uniquely determines a_n for all nonnegative integers n. We can find a formula for a_n from this information.

Some of the counting problems that cannot be solved using the techniques discussed in Chapter 5 can be solved by finding relationships, called recurrence relations, between the terms of a sequence, as was done in the problem involving bacteria. We will study a variety of counting problems that can be modeled using recurrence relations. We will develop methods in this section and in Section 7.2 for finding explicit formulae for the terms of sequences that satisfy certain types of recurrence relations.

Recurrence Relations

Assessment

In Chapter 4 we discussed how sequences can be defined recursively. Recall that a recursive definition of a sequence specifies one or more initial terms and a rule for determining subsequent terms from those that precede them. A rule of the latter sort (whether or not it is part of a

recursive definition) is called a **recurrence relation.** Such relations can be used in studying and solving counting problems.

DEFINITION 1 A *recurrence relation* for the sequence $\{a_n\}$ is an equation that expresses a_n in terms of one or more of the previous terms of the sequence, namely, $a_0, a_1, \ldots, a_{n-1}$, for all integers n with $n \geq n_0$, where n_0 is a nonnegative integer. A sequence is called a *solution* of a recurrence relation if its terms satisfy the recurrence relation.

There is an important connection between recursion and recurrence relations that we will exploit later in this chapter. A recursive algorithm provides the solution of a problem of size n in terms of the solutions of one or more instances of the same problem of smaller size. Consequently, when we analyze the complexity of a recursive algorithm, we obtain a recurrence relation that expresses the number of operations required to solve a problem of size n in terms of the number of operations required to solve the problem for one or more instances of smaller size.

EXAMPLE 1 Let $\{a_n\}$ be a sequence that satisfies the recurrence relation $a_n = a_{n-1} - a_{n-2}$ for $n = 2, 3, 4, \ldots$, and suppose that $a_0 = 3$ and $a_1 = 5$. What are a_2 and a_3?

Solution: We see from the recurrence relation that $a_2 = a_1 - a_0 = 5 - 3 = 2$ and $a_3 = a_2 - a_1 = 2 - 5 = -3$. We can find a_4, a_5, and each successive term in a similar way. ◀

EXAMPLE 2 Determine whether the sequence $\{a_n\}$, where $a_n = 3n$ for every nonnegative integer n, is a solution of the recurrence relation $a_n = 2a_{n-1} - a_{n-2}$ for $n = 2, 3, 4, \ldots$. Answer the same question where $a_n = 2^n$ and where $a_n = 5$.

Solution: Suppose that $a_n = 3n$ for every nonnegative integer n. Then, for $n \geq 2$, we see that $2a_{n-1} - a_{n-2} = 2[3(n-1)] - 3(n-2) = 3n = a_n$. Therefore, $\{a_n\}$, where $a_n = 3n$, is a solution of the recurrence relation.

Suppose that $a_n = 2^n$ for every nonnegative integer n. Note that $a_0 = 1$, $a_1 = 2$, and $a_2 = 4$. Because $2a_1 - a_0 = 2 \cdot 2 - 1 = 3 \neq a_2$, we see that $\{a_n\}$, where $a_n = 2^n$, is not a solution of the recurrence relation.

Suppose that $a_n = 5$ for every nonnegative integer n. Then for $n \geq 2$, we see that $a_n = 2a_{n-1} - a_{n-2} = 2 \cdot 5 - 5 = 5 = a_n$. Therefore, $\{a_n\}$, where $a_n = 5$, is a solution of the recurrence relation. ◀

The **initial conditions** for a sequence specify the terms that precede the first term where the recurrence relation takes effect. For instance, in Example 1, $a_0 = 3$ and $a_1 = 5$ are the initial conditions. The recurrence relation and initial conditions uniquely determine a sequence. This is the case because a recurrence relation, together with initial conditions, provide a recursive definition of the sequence. Any term of the sequence can be found from the initial conditions using the recurrence relation a sufficient number of times. However, there are better ways for computing the terms of certain classes of sequences defined by recurrence relations and initial conditions. We will discuss these methods in this section and in Section 7.2.

Modeling with Recurrence Relations

We can use recurrence relations to model a wide variety of problems, such as finding compound interest, counting rabbits on an island, determining the number of moves in the Tower of Hanoi puzzle, and counting bit strings with certain properties.

EXAMPLE 3 **Compound Interest** Suppose that a person deposits \$10,000 in a savings account at a bank yielding 11% per year with interest compounded annually. How much will be in the account after 30 years?

Extra Examples

Solution: To solve this problem, let P_n denote the amount in the account after n years. Because the amount in the account after n years equals the amount in the account after $n-1$ years plus interest for the nth year, we see that the sequence $\{P_n\}$ satisfies the recurrence relation

$$P_n = P_{n-1} + 0.11P_{n-1} = (1.11)P_{n-1}.$$

The initial condition is $P_0 = 10{,}000$.

We can use an iterative approach to find a formula for P_n. Note that

$$\begin{aligned}
P_1 &= (1.11)P_0 \\
P_2 &= (1.11)P_1 = (1.11)^2 P_0 \\
P_3 &= (1.11)P_2 = (1.11)^3 P_0 \\
&\vdots \\
P_n &= (1.11)P_{n-1} = (1.11)^n P_0.
\end{aligned}$$

When we insert the initial condition $P_0 = 10{,}000$, the formula $P_n = (1.11)^n 10{,}000$ is obtained.

We can use mathematical induction to establish its validity. That the formula is valid for $n = 0$ is a consequence of the initial condition. Now assume that $P_n = (1.11)^n 10{,}000$. Then, from the recurrence relation and the induction hypothesis,

$$P_{n+1} = (1.11)P_n = (1.11)(1.11)^n 10{,}000 = (1.11)^{n+1} 10{,}000.$$

This shows that the explicit formula for P_n is valid.

Inserting $n = 30$ into the formula $P_n = (1.11)^n 10{,}000$ shows that after 30 years the account contains

$$P_{30} = (1.11)^{30} 10{,}000 = \$228{,}922.97.$$ ◀

Example 4 shows how the population of rabbits on an island can be modeled using a recurrence relation.

EXAMPLE 4 **Rabbits and the Fibonacci Numbers** Consider this problem, which was originally posed by Leonardo Pisano, also known as Fibonacci, in the thirteenth century in his book *Liber abaci.* A young pair of rabbits (one of each sex) is placed on an island. A pair of rabbits does not breed until they are 2 months old. After they are 2 months old, each pair of rabbits produces another pair each month, as shown in Figure 1. Find a recurrence relation for the number of pairs of rabbits on the island after n months, assuming that no rabbits ever die.

Links

Solution: Denote by f_n the number of pairs of rabbits after n months. We will show that f_n, $n = 1, 2, 3, \ldots$, are the terms of the Fibonacci sequence.

The rabbit population can be modeled using a recurrence relation. At the end of the first month, the number of pairs of rabbits on the island is $f_1 = 1$. Because this pair does not breed during the second month, $f_2 = 1$ also. To find the number of pairs after n months, add the number on the island the previous month, f_{n-1}, and the number of newborn pairs, which equals f_{n-2}, because each newborn pair comes from a pair at least 2 months old.

Reproducing pairs (at least two months old)	Young pairs (less than two months old)	Month	Reproducing pairs	Young pairs	Total pairs
		1	0	1	1
		2	0	1	1
		3	1	1	2
		4	1	2	3
		5	2	3	5
		6	3	5	8

FIGURE 1 Rabbits on an Island.

Consequently, the sequence $\{f_n\}$ satisfies the recurrence relation

$$f_n = f_{n-1} + f_{n-2}$$

for $n \geq 3$ together with the initial conditions $f_1 = 1$ and $f_2 = 1$. Because this recurrence relation and the initial conditions uniquely determine this sequence, the number of pairs of rabbits on the island after n months is given by the nth Fibonacci number. ◀

Demo

Example 5 involves a famous puzzle.

EXAMPLE 5 **The Tower of Hanoi** A popular puzzle of the late nineteenth century invented by the French mathematician Édouard Lucas, called the Tower of Hanoi, consists of three pegs mounted on a board together with disks of different sizes. Initially these disks are placed on the first peg in order of size, with the largest on the bottom (as shown in Figure 2). The rules of the puzzle allow disks to be moved one at a time from one peg to another as long as a disk is never placed

Links

FIGURE 2 The Initial Position in the Tower of Hanoi.

FIGURE 3 An Intermediate Position in the Tower of Hanoi.

on top of a smaller disk. The goal of the puzzle is to have all the disks on the second peg in order of size, with the largest on the bottom.

Let H_n denote the number of moves needed to solve the Tower of Hanoi problem with n disks. Set up a recurrence relation for the sequence $\{H_n\}$.

Solution: Begin with n disks on peg 1. We can transfer the top $n-1$ disks, following the rules of the puzzle, to peg 3 using H_{n-1} moves (see Figure 3 for an illustration of the pegs and disks at this point). We keep the largest disk fixed during these moves. Then, we use one move to transfer the largest disk to the second peg. We can transfer the $n-1$ disks on peg 3 to peg 2 using H_{n-1} additional moves, placing them on top of the largest disk, which always stays fixed on the bottom of peg 2. Moreover, it is easy to see that the puzzle cannot be solved using fewer steps. This shows that

$$H_n = 2H_{n-1} + 1.$$

The initial condition is $H_1 = 1$, because one disk can be transferred from peg 1 to peg 2, according to the rules of the puzzle, in one move.

We can use an iterative approach to solve this recurrence relation. Note that

$$\begin{aligned}
H_n &= 2H_{n-1} + 1 \\
&= 2(2H_{n-2} + 1) + 1 = 2^2 H_{n-2} + 2 + 1 \\
&= 2^2(2H_{n-3} + 1) + 2 + 1 = 2^3 H_{n-3} + 2^2 + 2 + 1 \\
&\vdots \\
&= 2^{n-1} H_1 + 2^{n-2} + 2^{n-3} + \cdots + 2 + 1 \\
&= 2^{n-1} + 2^{n-2} + \cdots + 2 + 1 \\
&= 2^n - 1.
\end{aligned}$$

We have used the recurrence relation repeatedly to express H_n in terms of previous terms of the sequence. In the next to last equality, the initial condition $H_1 = 1$ has been used. The last equality is based on the formula for the sum of the terms of a geometric series, which can be found in Theorem 1 in Section 2.4.

The iterative approach has produced the solution to the recurrence relation $H_n = 2H_{n-1} + 1$ with the initial condition $H_1 = 1$. This formula can be proved using mathematical induction. This is left for the reader as an exercise at the end of the section.

A myth created to accompany the puzzle tells of a tower in Hanoi where monks are transferring 64 gold disks from one peg to another, according to the rules of the puzzle. The myth

says that the world will end when they finish the puzzle. How long after the monks started will the world end if the monks take one second to move a disk?

From the explicit formula, the monks require

$$2^{64} - 1 = 18{,}446{,}744{,}073{,}709{,}551{,}615$$

moves to transfer the disks. Making one move per second, it will take them more than 500 billion years to complete the transfer, so the world should survive a while longer than it already has. ◀

Links

Remark: Many people have studied variations of the original Tower of Hanoi puzzle discussed in Example 5. Some variations use more pegs, some allow disks to be of the same size, and some restrict the types of allowable disk moves. One of the oldest and most interesting variations is the **Reve's puzzle,*** proposed in 1907 by Henry Dudeney in his book *The Canterbury Puzzles*. The Reve's puzzle involves pilgrims challenged by the Reve to move a stack of cheeses of varying sizes from the first of four stools to another stool without ever placing a cheese on one of smaller diameter. The Reve's puzzle, expressed in terms of pegs and disks, follows the same rules as the Tower of Hanoi puzzle, except that four pegs are used. You may find it surprising that no one has been able to establish the minimum number of moves required to solve this puzzle for n disks. However, there is a conjecture, now more than 50 years old, that the minimum number of moves required equals the number of moves used by an algorithm invented by Frame and Stewart in 1939. (See Exercises 54–61 at the end of this section and [St94] for more information.)

Example 6 illustrates how recurrence relations can be used to count bit strings of a specified length that have a certain property.

EXAMPLE 6 Find a recurrence relation and give initial conditions for the number of bit strings of length n that do not have two consecutive 0s. How many such bit strings are there of length five?

Solution: Let a_n denote the number of bit strings of length n that do not have two consecutive 0s. To obtain a recurrence relation for $\{a_n\}$, note that by the sum rule, the number of bit strings of length n that do not have two consecutive 0s equals the number of such bit strings ending with a 0 plus the number of such bit strings ending with a 1. We will assume that $n \geq 3$, so that the bit string has at least three bits.

The bit strings of length n ending with 1 that do not have two consecutive 0s are precisely the bit strings of length $n - 1$ with no two consecutive 0s with a 1 added at the end. Consequently, there are a_{n-1} such bit strings.

Bit strings of length n ending with a 0 that do not have two consecutive 0s must have 1 as their $(n - 1)$st bit; otherwise they would end with a pair of 0s. It follows that the bit strings of length n ending with a 0 that have no two consecutive 0s are precisely the bit strings of length $n - 2$ with no two consecutive 0s with 10 added at the end. Consequently, there are a_{n-2} such bit strings.

We conclude, as illustrated in Figure 4, that

$$a_n = a_{n-1} + a_{n-2}$$

for $n \geq 3$.

The initial conditions are $a_1 = 2$, because both bit strings of length one, 0 and 1 do not have consecutive 0s, and $a_2 = 3$, because the valid bit strings of length two are 01, 10, and 11. To

*Reve, more commonly spelled *reeve*, is an archaic word for *governor*.

FIGURE 4 Counting Bit Strings of Length n with No Two Consecutive 0s.

obtain a_5, we use the recurrence relation three times to find that

$$\begin{aligned} a_3 &= a_2 + a_1 = 3 + 2 = 5, \\ a_4 &= a_3 + a_2 = 5 + 3 = 8, \\ a_5 &= a_4 + a_3 = 8 + 5 = 13. \end{aligned}$$

◀

Remark: Note that $\{a_n\}$ satisfies the same recurrence relation as the Fibonacci sequence. Because $a_1 = f_3$ and $a_2 = f_4$ it follows that $a_n = f_{n+2}$.

Example 7 shows how a recurrence relation can be used to model the number of codewords that are allowable using certain validity checks.

EXAMPLE 7 **Codeword Enumeration** A computer system considers a string of decimal digits a valid codeword if it contains an even number of 0 digits. For instance, 1230407869 is valid, whereas 120987045608 is not valid. Let a_n be the number of valid n-digit codewords. Find a recurrence relation for a_n.

Solution: Note that $a_1 = 9$ because there are 10 one-digit strings, and only one, namely, the string 0, is not valid. A recurrence relation can be derived for this sequence by considering how a valid n-digit string can be obtained from strings of $n - 1$ digits. There are two ways to form a valid string with n digits from a string with one fewer digit.

First, a valid string of n digits can be obtained by appending a valid string of $n - 1$ digits with a digit other than 0. This appending can be done in nine ways. Hence, a valid string with n digits can be formed in this manner in $9a_{n-1}$ ways.

Second, a valid string of n digits can be obtained by appending a 0 to a string of length $n - 1$ that is not valid. (This produces a string with an even number of 0 digits because the invalid string of length $n - 1$ has an odd number of 0 digits.) The number of ways that this can be done equals the number of invalid $(n - 1)$-digit strings. Because there are 10^{n-1} strings of length $n - 1$, and a_{n-1} are valid, there are $10^{n-1} - a_{n-1}$ valid n-digit strings obtained by appending an invalid string of length $n - 1$ with a 0.

Because all valid strings of length n are produced in one of these two ways, it follows that there are

$$\begin{aligned} a_n &= 9a_{n-1} + (10^{n-1} - a_{n-1}) \\ &= 8a_{n-1} + 10^{n-1} \end{aligned}$$

valid strings of length n. ◀

Example 8 establishes a recurrence relation that appears in many different contexts.

EXAMPLE 8 Find a recurrence relation for C_n, the number of ways to parenthesize the product of $n+1$ numbers, $x_0 \cdot x_1 \cdot x_2 \cdot \cdots \cdot x_n$, to specify the order of multiplication. For example, $C_3 = 5$ because there are five ways to parenthesize $x_0 \cdot x_1 \cdot x_2 \cdot x_3$ to determine the order of multiplication:

$$((x_0 \cdot x_1) \cdot x_2) \cdot x_3 \qquad (x_0 \cdot (x_1 \cdot x_2)) \cdot x_3 \qquad (x_0 \cdot x_1) \cdot (x_2 \cdot x_3)$$
$$x_0 \cdot ((x_1 \cdot x_2) \cdot x_3) \qquad x_0 \cdot (x_1 \cdot (x_2 \cdot x_3)).$$

Solution: To develop a recurrence relation for C_n, we note that however we insert parentheses in the product $x_0 \cdot x_1 \cdot x_2 \cdot \cdots \cdot x_n$, one "$\cdot$" operator remains outside all parentheses, namely, the operator for the final multiplication to be performed. [For example, in $(x_0 \cdot (x_1 \cdot x_2)) \cdot x_3$, it is the final "$\cdot$", while in $(x_0 \cdot x_1) \cdot (x_2 \cdot x_3)$ it is the second "$\cdot$".] This final operator appears between two of the $n+1$ numbers, say, x_k and x_{k+1}. There are $C_k C_{n-k-1}$ ways to insert parentheses to determine the order of the $n+1$ numbers to be multiplied when the final operator appears between x_k and x_{k+1}, because there are C_k ways to insert parentheses in the product $x_0 \cdot x_1 \cdot \cdots \cdot x_k$ to determine the order in which these $k+1$ numbers are to be multiplied and C_{n-k-1} ways to insert parentheses in the product $x_{k+1} \cdot x_{k+2} \cdot \cdots \cdot x_n$ to determine the order in which these $n-k$ numbers are to be multiplied. Because this final operator can appear between any two of the $n+1$ numbers, it follows that

$$\begin{aligned} C_n &= C_0 C_{n-1} + C_1 C_{n-2} + \cdots + C_{n-2} C_1 + C_{n-1} C_0 \\ &= \sum_{k=0}^{n-1} C_k C_{n-k-1}. \end{aligned}$$

Note that the initial conditions are $C_0 = 1$ and $C_1 = 1$. This recurrence relation can be solved using the method of generating functions, which will be discussed in Section 7.4. It can be shown that $C_n = C(2n, n)/(n+1)$. (See Exercise 41 at the end of that section.) ◀

Links The sequence $\{C_n\}$ is the sequence of **Catalan numbers.** This sequence appears as the solution of many different counting problems besides the one considered here (see the chapter on Catalan numbers in [MiRo91] or [Ro84a] for details).

Exercises

1. Find the first five terms of the sequence defined by each of these recurrence relations and initial conditions.
   a) $a_n = 6a_{n-1}, a_0 = 2$
   b) $a_n = a_{n-1}^2, a_1 = 2$
   c) $a_n = a_{n-1} + 3a_{n-2}, a_0 = 1, a_1 = 2$
   d) $a_n = na_{n-1} + n^2 a_{n-2}, a_0 = 1, a_1 = 1$
   e) $a_n = a_{n-1} + a_{n-3}, a_0 = 1, a_1 = 2, a_2 = 0$
2. Find the first six terms of the sequence defined by each of these recurrence relations and initial conditions.
   a) $a_n = -2a_{n-1}, a_0 = -1$
   b) $a_n = a_{n-1} - a_{n-2}, a_0 = 2, a_1 = -1$
   c) $a_n = 3a_{n-1}^2, a_0 = 1$
   d) $a_n = na_{n-1} + a_{n-2}^2, a_0 = -1, a_1 = 0$
   e) $a_n = a_{n-1} - a_{n-2} + a_{n-3}, a_0 = 1, a_1 = 1, a_2 = 2$
3. Let $a_n = 2^n + 5 \cdot 3^n$ for $n = 0, 1, 2, \ldots$.
   a) Find a_0, a_1, a_2, a_3, and a_4.
   b) Show that $a_2 = 5a_1 - 6a_0$, $a_3 = 5a_2 - 6a_1$, and $a_4 = 5a_3 - 6a_2$.
   c) Show that $a_n = 5a_{n-1} - 6a_{n-2}$ for all integers n with $n \geq 2$.
4. Show that the sequence $\{a_n\}$ is a solution of the recurrence relation $a_n = -3a_{n-1} + 4a_{n-2}$ if
   a) $a_n = 0$. b) $a_n = 1$.
   c) $a_n = (-4)^n$. d) $a_n = 2(-4)^n + 3$.
5. Is the sequence $\{a_n\}$ a solution of the recurrence relation $a_n = 8a_{n-1} - 16a_{n-2}$ if
   a) $a_n = 0$? b) $a_n = 1$?
   c) $a_n = 2^n$? d) $a_n = 4^n$?
   e) $a_n = n4^n$? f) $a_n = 2 \cdot 4^n + 3n4^n$?
   g) $a_n = (-4)^n$? h) $a_n = n^2 4^n$?
6. For each of these sequences find a recurrence relation satisfied by this sequence. (The answers are not unique

because there are infinitely many different recurrence relations satisfied by any sequence.)

a) $a_n = 3$ b) $a_n = 2n$
c) $a_n = 2n + 3$ d) $a_n = 5^n$
e) $a_n = n^2$ f) $a_n = n^2 + n$
g) $a_n = n + (-1)^n$ h) $a_n = n!$

7. Show that the sequence $\{a_n\}$ is a solution of the recurrence relation $a_n = a_{n-1} + 2a_{n-2} + 2n - 9$ if

a) $a_n = -n + 2$.
b) $a_n = 5(-1)^n - n + 2$.
c) $a_n = 3(-1)^n + 2^n - n + 2$.
d) $a_n = 7 \cdot 2^n - n + 2$.

8. Find the solution to each of these recurrence relations with the given initial conditions. Use an iterative approach such as that used in Example 5.

a) $a_n = -a_{n-1}, a_0 = 5$
b) $a_n = a_{n-1} + 3, a_0 = 1$
c) $a_n = a_{n-1} - n, a_0 = 4$
d) $a_n = 2a_{n-1} - 3, a_0 = -1$
e) $a_n = (n + 1)a_{n-1}, a_0 = 2$
f) $a_n = 2na_{n-1}, a_0 = 3$
g) $a_n = -a_{n-1} + n - 1, a_0 = 7$

9. Find the solution to each of these recurrence relations and initial conditions. Use an iterative approach such as that used in Example 5.

a) $a_n = 3a_{n-1}, a_0 = 2$
b) $a_n = a_{n-1} + 2, a_0 = 3$
c) $a_n = a_{n-1} + n, a_0 = 1$
d) $a_n = a_{n-1} + 2n + 3, a_0 = 4$
e) $a_n = 2a_{n-1} - 1, a_0 = 1$
f) $a_n = 3a_{n-1} + 1, a_0 = 1$
g) $a_n = na_{n-1}, a_0 = 5$
h) $a_n = 2na_{n-1}, a_0 = 1$

10. A person deposits \$1000 in an account that yields 9% interest compounded anually.

a) Set up a recurrence relation for the amount in the account at the end of n years.
b) Find an explicit formula for the amount in the account at the end of n years.
c) How much money will the account contain after 100 years?

11. Suppose that the number of bacteria in a colony triples every hour.

a) Set up a recurrence relation for the number of bacteria after n hours have elapsed.
b) If 100 bacteria are used to begin a new colony, how many bacteria will be in the colony in 10 hours?

12. Assume that the population of the world in 2002 was 6.2 billion and is growing at the rate of 1.3% a year.

a) Set up a recurrence relation for the population of the world n years after 2002.
b) Find an explicit formula for the population of the world n years after 2002.
c) What will the population of the world be in 2022?

13. A factory makes custom sports cars at an increasing rate. In the first month only one car is made, in the second month two cars are made, and so on, with n cars made in the nth month.

a) Set up a recurrence relation for the number of cars produced in the first n months by this factory.
b) How many cars are produced in the first year?
c) Find an explicit formula for the number of cars produced in the first n months by this factory.

14. An employee joined a company in 1999 with a starting salary of \$50,000. Every year this employee receives a raise of \$1000 plus 5% of the salary of the previous year.

a) Set up a recurrence relation for the salary of this employee n years after 1999.
b) What will the salary of this employee be in 2007?
c) Find an explicit formula for the salary of this employee n years after 1999.

15. Find a recurrence relation for the balance $B(k)$ owed at the end of k months on a loan of \$5000 at a rate of 7% if a payment of \$100 is made each month. [*Hint:* Express $B(k)$ in terms of $B(k-1)$; the monthly interest is $(0.07/12)B(k-1)$.]

16. a) Find a recurrence relation for the balance $B(k)$ owed at the end of k months on a loan at a rate of r if a payment P is made on the loan each month. [*Hint:* Express $B(k)$ in terms of $B(k-1)$ and note that the monthly interest rate is $r/12$.]
b) Determine what the monthly payment P should be so that the loan is paid off after T months.

17. Use mathematical induction to verify the formula derived in Example 5 for the number of moves required to complete the Tower of Hanoi puzzle.

18. a) Find a recurrence relation for the number of permutations of a set with n elements.
b) Use this recurrence relation to find the number of permutations of a set with n elements using iteration.

19. A vending machine dispensing books of stamps accepts only dollar coins, \$1 bills, and \$5 bills.

a) Find a recurrence relation for the number of ways to deposit n dollars in the vending machine, where the order in which the coins and bills are deposited matters.
b) What are the initial conditions?
c) How many ways are there to deposit \$10 for a book of stamps?

20. A country uses as currency coins with values of 1 peso, 2 pesos, 5 pesos, and 10 pesos and bills with values of 5 pesos, 10 pesos, 20 pesos, 50 pesos, and 100 pesos. Find a recurrence relation for the number of ways to pay a bill of n pesos if the order in which the coins and bills are paid matters.

21. How many ways are there to pay a bill of 17 pesos using the currency described in Exercise 20, where the order in which coins and bills are paid matters?

*22. a) Find a recurrence relation for the number of strictly increasing sequences of positive integers that have 1 as their first term and n as their last term, where n is a positive integer. That is, sequences $a_1, a_2, \ldots, a_k$, where $a_1 = 1$, $a_k = n$, and $a_j < a_{j+1}$ for $j = 1, 2, \ldots, k-1$.
b) What are the initial conditions?
c) How many sequences of the type described in (a) are there when n is a positive integer with $n \geq 2$?

23. a) Find a recurrence relation for the number of bit strings of length n that contain a pair of consecutive 0s.
b) What are the initial conditions?
c) How many bit strings of length seven contain two consecutive 0s?

24. a) Find a recurrence relation for the number of bit strings of length n that contain three consecutive 0s.
b) What are the initial conditions?
c) How many bit strings of length seven contain three consecutive 0s?

25. a) Find a recurrence relation for the number of bit strings of length n that do not contain three consecutive 0s.
b) What are the initial conditions?
c) How many bit strings of length seven do not contain three consecutive 0s?

26. a) Find a recurrence relation for the number of bit strings that contain the string 01.
b) What are the initial conditions?
c) How many bit strings of length seven contain the string 01?

27. a) Find a recurrence relation for the number of ways to climb n stairs if the person climbing the stairs can take one stair or two stairs at a time.
b) What are the initial conditions?
c) How many ways can this person climb a flight of eight stairs?

28. a) Find a recurrence relation for the number of ways to climb n stairs if the person climbing the stairs can take one, two, or three stairs at a time.
b) What are the initial conditions?
c) How many ways can this person climb a flight of eight stairs?

A string that contains only 0s, 1s, and 2s is called a **ternary string.**

29. a) Find a recurrence relation for the number of ternary strings that do not contain two consecutive 0s.
b) What are the initial conditions?
c) How many ternary strings of length six do not contain two consecutive 0s?

30. a) Find a recurrence relation for the number of ternary strings that contain two consecutive 0s.
b) What are the initial conditions?
c) How many ternary strings of length six contain two consecutive 0s?

31. a) Find a recurrence relation for the number of ternary strings that do not contain two consecutive 0s or two consecutive 1s.
b) What are the initial conditions?
c) How many ternary strings of length six do not contain two consecutive 0s or two consecutive 1s?

*32. a) Find a recurrence relation for the number of ternary strings that contain either two consecutive 0s or two consecutive 1s.
b) What are the initial conditions?
c) How many ternary strings of length six contain two consecutive 0s or two consecutive 1s?

*33. a) Find a recurrence relation for the number of ternary strings that do not contain consecutive symbols that are the same.
b) What are the initial conditions?
c) How many ternary strings of length six do not contain consecutive symbols that are the same?

**34. a) Find a recurrence relation for the number of ternary strings that contain two consecutive symbols that are the same.
b) What are the initial conditions?
c) How many ternary strings of length six contain consecutive symbols that are the same?

35. Messages are transmitted over a communications channel using two signals. The transmittal of one signal requires 1 microsecond, and the transmittal of the other signal requires 2 microseconds.
a) Find a recurrence relation for the number of different messages consisting of sequences of these two signals, where each signal in the message is immediately followed by the next signal, that can be sent in n microseconds.
b) What are the initial conditions?
c) How many different messages can be sent in 10 microseconds using these two signals?

36. A bus driver pays all tolls, using only nickels and dimes, by throwing one coin at a time into the mechanical toll collector.
a) Find a recurrence relation for the number of different ways the bus driver can pay a toll of n cents (where the order in which the coins are used matters).
b) In how many different ways can the driver pay a toll of 45 cents?

37. a) Find the recurrence relation satisfied by R_n, where R_n is the number of regions that a plane is divided into by n lines, if no two of the lines are parallel and no three of the lines go through the same point.
b) Find R_n using iteration.

*38. a) Find the recurrence relation satisfied by R_n, where R_n is the number of regions into which the surface of a sphere is divided by n great circles (which are the intersections of the sphere and planes passing through the center of the sphere), if no three of the great circles go through the same point.
b) Find R_n using iteration.

*39. a) Find the recurrence relation satisfied by S_n, where S_n is the number of regions into which three-dimensional space is divided by n planes if every three of the planes

meet in one point, but no four of the planes go through the same point.

b) Find S_n using iteration.

40. Find a recurrence relation for the number of bit sequences of length n with an even number of 0s.

41. How many bit sequences of length seven contain an even number of 0s?

42. a) Find a recurrence relation for the number of ways to completely cover a $2 \times n$ checkerboard with 1×2 dominoes. [*Hint:* Consider separately the coverings where the position in the top right corner of the checkerboard is covered by a domino positioned horizontally and where it is covered by a domino positioned vertically.]

b) What are the initial conditions for the recurrence relation in part (a)?

c) How many ways are there to completely cover a 2×17 checkerboard with 1×2 dominoes?

43. a) Find a recurrence relation for the number of ways to lay out a walkway with slate tiles if the tiles are red, green, or gray, so that no two red tiles are adjacent and tiles of the same color are considered indistinguishable.

b) What are the initial conditions for the recurrence relation in part (a)?

c) How many ways are there to lay out a path of seven tiles as described in part (a)?

44. Show that the Fibonacci numbers satisfy the recurrence relation $f_n = 5f_{n-4} + 3f_{n-5}$ for $n = 5, 6, 7, \ldots$, together with the initial conditions $f_0 = 0, f_1 = 1, f_2 = 1, f_3 = 2$, and $f_4 = 3$. Use this recurrence relation to show that f_{5n} is divisible by 5, for $n = 1, 2, 3, \ldots$.

k45. Let $S(m, n)$ denote the number of onto functions from a set with m elements to a set with n elements. Show that $S(m, n)$ satisfies the recurrence relation

$$S(m, n) = n^m - \sum_{k=1}^{n-1} C(n, k)S(m, k)$$

whenever $m \geq n$ and $n > 1$, with the initial condition $S(m, 1) = 1$.

46. a) Write out all the ways the product $x_0 \cdot x_1 \cdot x_2 \cdot x_3 \cdot x_4$ can be parenthesized to determine the order of multiplication.

b) Use the recurrence relation developed in Example 8 to calculate C_4, the number of ways to parenthesize the product of five numbers so as to determine the order of multiplication. Verify that you listed the correct number of ways in part (a).

c) Check your result in part (b) by finding C_4, using the closed formula for C_n mentioned in the solution of Example 8.

47. a) Use the recurrence relation developed in Example 8 to determine C_5, the number of ways to parenthesize the product of six numbers so as to determine the order of multiplication.

b) Check your result with the closed formula for C_5 mentioned in the solution of Example 8.

***48.** In the Tower of Hanoi puzzle, suppose our goal is to transfer all n disks from peg 1 to peg 3, but we cannot move a disk directly between pegs 1 and 3. Each move of a disk must be a move involving peg 2. As usual, we cannot place a disk on top of a smaller disk.

a) Find a recurrence relation for the number of moves required to solve the puzzle for n disks with this added restriction.

b) Solve this recurrence relation to find a formula for the number of moves required to solve the puzzle for n disks.

c) How many different arrangements are there of the n disks on three pegs so that no disk is on top of a smaller disk?

d) Show that every allowable arrangement of the n disks occurs in the solution of this variation of the puzzle.

Exercises 49–53 deal with a variation of the **Josephus problem** described by Graham, Knuth, and Patashnik in [GrKnPa94]. This problem is based on an account by the historian Flavius Josephus, who was part of a band of 41 Jewish rebels trapped in a cave by the Romans during the Jewish-Roman war of the first century. The rebels preferred suicide to capture; they decided to form a circle and to repeatedly count off around the circle, killing every third rebel left alive. However, Josephus and another rebel did not want to be killed this way; they determined the positions where they should stand to be the last two rebels remaining alive. The variation we consider begins with n people, numbered 1 to n, standing around a circle. In each stage, every second person still left alive is eliminated until only one survives. We denote the number of the survivor by $J(n)$.

49. Determine the value of $J(n)$ for each integer n with $1 \leq n \leq 16$.

50. Use the values you found in Exercise 49 to conjecture a formula for $J(n)$. [*Hint:* Write $n = 2^m + k$, where m is a nonnegative integer and k is a nonnegative integer less than 2^m.]

51. Show that $J(n)$ satisfies the recurrence relation $J(2n) = 2J(n) - 1$ and $J(2n + 1) = 2J(n) + 1$, for $n \geq 1$, and $J(1) = 1$.

52. Use mathematical induction to prove the formula you conjectured in Exercise 50, making use of the recurrence relation from Exercise 51.

53. Determine $J(100)$, $J(1000)$, and $J(10{,}000)$ from your formula for $J(n)$.

Exercises 54–61 involve the Reve's puzzle, the variation of the Tower of Hanoi puzzle with four pegs and n disks. Before presenting these exercises, we describe the Frame–Stewart algorithm for moving the disks from peg 1 to peg 4 so that no disk is ever on top of a smaller one. This algorithm, given the number of disks n as input, depends on a choice of an integer k with $1 \leq k \leq n$. When there is only one disk, move it from peg 1 to peg 4 and stop. For $n > 1$, the algorithm proceeds

recursively, using these three steps. Recursively move the stack of the $n-k$ smallest disks from peg 1 to peg 2, using all four pegs. Next move the stack of the k largest disks from peg 1 to peg 4, using the three-peg algorithm from the Tower of Hanoi puzzle without using the peg holding the $n-k$ smallest disks. Finally, recursively move the smallest $n-k$ disks to peg 4, using all four pegs. Frame and Stewart showed that to produce the fewest moves using their algorithm, k should be chosen to be the smallest integer such that n does not exceed $t_k = k(k+1)/2$, the kth triangular number, that is, $t_{k-1} < n \leq t_k$. The unsettled conjecture, known as **Frame's conjecture,** is that this algorithm uses the fewest number of moves required to solve the puzzle, no matter how the disks are moved.

54. Show that the Reve's puzzle with three disks can be solved using five, and no fewer, moves.

55. Show that the Reve's puzzle with four disks can be solved using nine, and no fewer, moves.

56. Describe the moves made by the Frame–Stewart algorithm, with k chosen so that the fewest moves are required, for

a) 5 disks. **b)** 6 disks. **c)** 7 disks. **d)** 8 disks.

57. Show that if $R(n)$ is the number of moves used by the Frame–Stewart algorithm to solve the Reve's puzzle with n disks, where k is chosen to be the smallest integer with $n \leq k(k+1)/2$, then $R(n)$ satisfies the recurrence relation $R(n) = 2R(n-k) + 2^k - 1$, with $R(0) = 0$ and $R(1) = 1$.

58. Show that if k is as chosen in Exercise 57, then $R(n) - R(n-1) = 2^{k-1}$.

***59.** Show that if k is as chosen in Exercise 57, then $R(n) = \sum_{i=1}^{k} i2^{i-1} - (t_k - n)2^{k-1}$.

***60.** Use Exercise 59 to give an upper bound on the number of moves required to solve the Reve's puzzle for all integers n with $1 \leq n \leq 25$.

***61.** Show that $R(n)$ is $O(\sqrt{n}2^{\sqrt{2n}})$.

Let $\{a_n\}$ be a sequence of real numbers. The **backward differences** of this sequence are defined recursively as shown next. The **first difference** ∇a_n is

$$\nabla a_n = a_n - a_{n-1}.$$

The **$(k+1)$st difference** $\nabla^{k+1} a_n$ is obtained from $\nabla^k a_n$ by

$$\nabla^{k+1} a_n = \nabla^k a_n - \nabla^k a_{n-1}.$$

62. Find ∇a_n for the sequence $\{a_n\}$, where

a) $a_n = 4$. **b)** $a_n = 2n$.
c) $a_n = n^2$. **d)** $a_n = 2^n$.

63. Find $\nabla^2 a_n$ for the sequences in Exercise 62.

64. Show that $a_{n-1} = a_n - \nabla a_n$.

65. Show that $a_{n-2} = a_n - 2\nabla a_n + \nabla^2 a_n$.

***66.** Prove that a_{n-k} can be expressed in terms of a_n, ∇a_n, $\nabla^2 a_n, \ldots, \nabla^k a_n$.

67. Express the recurrence relation $a_n = a_{n-1} + a_{n-2}$ in terms of a_n, ∇a_n, and $\nabla^2 a_n$.

68. Show that any recurrence relation for the sequence $\{a_n\}$ can be written in terms of a_n, ∇a_n, $\nabla^2 a_n, \ldots$. The resulting equation involving the sequences and its differences is called a **difference equation.**

7.2 Solving Linear Recurrence Relations

Introduction

Links

A wide variety of recurrence relations occur in models. Some of these recurrence relations can be solved using iteration or some other ad hoc technique. However, one important class of recurrence relations can be explicitly solved in a systematic way. These are recurrence relations that express the terms of a sequence as linear combinations of previous terms.

DEFINITION 1 A *linear homogeneous recurrence relation of degree k with constant coefficients* is a recurrence relation of the form

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_k a_{n-k},$$

where $c_1, c_2, \ldots, c_k$ are real numbers, and $c_k \neq 0$.

The recurrence relation in the definition is **linear** because the right-hand side is a sum of previous terms of the sequence each multiplied by a function of n. The recurrence relation is

homogeneous because no terms occur that are not multiples of the a_js. The coefficients of the terms of the sequence are all **constants,** rather than functions that depend on n. The **degree** is k because a_n is expressed in terms of the previous k terms of the sequence.

A consequence of the second principle of mathematical induction is that a sequence satisfying the recurrence relation in the definition is uniquely determined by this recurrence relation and the k initial conditions

$$a_0 = C_0, a_1 = C_1, \ldots, a_{k-1} = C_{k-1}.$$

EXAMPLE 1 The recurrence relation $P_n = (1.11)P_{n-1}$ is a linear homogeneous recurrence relation of degree one. The recurrence relation $f_n = f_{n-1} + f_{n-2}$ is a linear homogeneous recurrence relation of degree two. The recurrence relation $a_n = a_{n-5}$ is a linear homogeneous recurrence relation of degree five. ◀

Example 2 presents some examples of recurrence relations that are not linear homogeneous recurrence relations with constant coefficients.

EXAMPLE 2 The recurrence relation $a_n = a_{n-1} + a_{n-2}^2$ is not linear. The recurrence relation $H_n = 2H_{n-1} + 1$ is not homogeneous. The recurrence relation $B_n = nB_{n-1}$ does not have constant coefficients. ◀

Linear homogeneous recurrence relations are studied for two reasons. First, they often occur in modeling of problems. Second, they can be systematically solved.

Solving Linear Homogeneous Recurrence Relations with Constant Coefficients

The basic approach for solving linear homogeneous recurrence relations is to look for solutions of the form $a_n = r^n$, where r is a constant. Note that $a_n = r^n$ is a solution of the recurrence relation $a_n = c_1a_{n-1} + c_2a_{n-2} + \cdots + c_ka_{n-k}$ if and only if

$$r^n = c_1r^{n-1} + c_2r^{n-2} + \cdots + c_kr^{n-k}.$$

When both sides of this equation are divided by r^{n-k} and the right-hand side is subtracted from the left, we obtain the equation

$$r^k - c_1r^{k-1} - c_2r^{k-2} - \cdots - c_{k-1}r - c_k = 0.$$

Consequently, the sequence $\{a_n\}$ with $a_n = r^n$ is a solution if and only if r is a solution of this last equation, which is called the **characteristic equation** of the recurrence relation. The solutions of this equation are called the **characteristic roots** of the recurrence relation. As we will see, these characteristic roots can be used to give an explicit formula for all the solutions of the recurrence relation.

We will first develop results that deal with linear homogeneous recurrence relations with constant coefficients of degree two. Then corresponding general results when the degree may be greater than two will be stated. Because the proofs needed to establish the results in the general case are more complicated, they will not be given in the text.

We now turn our attention to linear homogeneous recurrence relations of degree two. First, consider the case when there are two distinct characteristic roots.

THEOREM 1 Let c_1 and c_2 be real numbers. Suppose that $r^2 - c_1 r - c_2 = 0$ has two distinct roots r_1 and r_2. Then the sequence $\{a_n\}$ is a solution of the recurrence relation $a_n = c_1 a_{n-1} + c_2 a_{n-2}$ if and only if $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ for $n = 0, 1, 2, \ldots$, where α_1 and α_2 are constants.

Proof: We must do two things to prove the theorem. First, it must be shown that if r_1 and r_2 are the roots of the characteristic equation, and α_1 and α_2 are constants, then the sequence $\{a_n\}$ with $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ is a solution of the recurrence relation. Second, it must be shown that if the sequence $\{a_n\}$ is a solution, then $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ for some constants α_1 and α_2.

Now we will show that if $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$, then the sequence $\{a_n\}$ is a solution of the recurrence relation. Because r_1 and r_2 are roots of $r^2 - c_1 r - c_2 = 0$, it follows that $r_1^2 = c_1 r_1 + c_2$, $r_2^2 = c_1 r_2 + c_2$.

From these equations, we see that

$$\begin{aligned} c_1 a_{n-1} + c_2 a_{n-2} &= c_1(\alpha_1 r_1^{n-1} + \alpha_2 r_2^{n-1}) + c_2(\alpha_1 r_1^{n-2} + \alpha_2 r_2^{n-2}) \\ &= \alpha_1 r_1^{n-2}(c_1 r_1 + c_2) + \alpha_2 r_2^{n-2}(c_1 r_2 + c_2) \\ &= \alpha_1 r_1^{n-2} r_1^2 + \alpha_2 r_2^{n-2} r_2^2 \\ &= \alpha_1 r_1^n + \alpha_2 r_2^n. \\ &= a_n. \end{aligned}$$

This shows that the sequence $\{a_n\}$ with $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ is a solution of the recurrence relation.

To show that every solution $\{a_n\}$ of the recurrence relation $a_n = c_1 a_{n-1} + c_2 a_{n-2}$ has $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ for $n = 0, 1, 2, \ldots$, for some constants α_1 and α_2, suppose that $\{a_n\}$ is a solution of the recurrence relation, and the initial conditions $a_0 = C_0$ and $a_1 = C_1$ hold. It will be shown that there are constants α_1 and α_2 such that the sequence $\{a_n\}$ with $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ satisfies these same initial conditions. This requires that

$$\begin{aligned} a_0 &= C_0 = \alpha_1 + \alpha_2, \\ a_1 &= C_1 = \alpha_1 r_1 + \alpha_2 r_2. \end{aligned}$$

We can solve these two equations for α_1 and α_2. From the first equation it follows that $\alpha_2 = C_0 - \alpha_1$. Inserting this expression into the second equation gives

$$C_1 = \alpha_1 r_1 + (C_0 - \alpha_1) r_2.$$

Hence,

$$C_1 = \alpha_1 (r_1 - r_2) + C_0 r_2.$$

This shows that

$$\alpha_1 = \frac{C_1 - C_0 r_2}{r_1 - r_2}$$

and

$$\alpha_2 = C_0 - \alpha_1 = C_0 - \frac{C_1 - C_0 r_2}{r_1 - r_2} = \frac{C_0 r_1 - C_1}{r_1 - r_2},$$

where these expressions for α_1 and α_2 depend on the fact that $r_1 \neq r_2$. (When $r_1 = r_2$, this theorem is not true.) Hence, with these values for α_1 and α_2, the sequence $\{a_n\}$ with $\alpha_1 r_1^n + \alpha_2 r_2^n$ satisfies the two initial conditions.

We know that $\{a_n\}$ and $\{\alpha_1 r_1^n + \alpha_2 r_2^n\}$ are both solutions of the recurrence relation $a_n = c_1 a_{n-1} + c_2 a_{n-2}$ and both satisfy the initial conditions when $n = 0$ and $n = 1$. Because there is a unique solution of a linear homogeneous recurrence relation of degree two with two initial conditions, it follows that the two solutions are the same, that is, $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$ for all nonnegative integers n. We have completed the proof by showing that a solution of the linear homogeneous recurrence relation with constant coefficients of degree two must be of the form $a_n = \alpha_1 r_1^n + \alpha_2 r_2^n$, where α_1 and α_2 are constants. ◁

The characteristic roots of a linear homogeneous recurrence relation with constant coefficients may be complex numbers. Theorem 1 (and also subsequent theorems in this section) still applies in this case. Recurrence relations with complex characteristic roots will not be discussed in the text. Readers familiar with complex numbers may wish to solve Exercises 38 and 39 at the end of this section.

Examples 3 and 4 show how to use Theorem 1 to solve recurrence relations.

EXAMPLE 3 What is the solution of the recurrence relation

$$a_n = a_{n-1} + 2a_{n-2}$$

with $a_0 = 2$ and $a_1 = 7$?

Extra Examples

Solution: Theorem 1 can be used to solve this problem. The characteristic equation of the recurrence relation is $r^2 - r - 2 = 0$. Its roots are $r = 2$ and $r = -1$. Hence, the sequence $\{a_n\}$ is a solution to the recurrence relation if and only if

$$a_n = \alpha_1 2^n + \alpha_2(-1)^n,$$

for some constants α_1 and α_2. From the initial conditions, it follows that

$$a_0 = 2 = \alpha_1 + \alpha_2,$$
$$a_1 = 7 = \alpha_1 \cdot 2 + \alpha_2 \cdot (-1).$$

Solving these two equations shows that $\alpha_1 = 3$ and $\alpha_2 = -1$. Hence, the solution to the recurrence relation and initial conditions is the sequence $\{a_n\}$ with

$$a_n = 3 \cdot 2^n - (-1)^n.$$ ◀

EXAMPLE 4 Find an explicit formula for the Fibonacci numbers.

Solution: Recall that the sequence of Fibonacci numbers satisfies the recurrence relation $f_n = f_{n-1} + f_{n-2}$ and also satisfies the initial conditions $f_0 = 0$ and $f_1 = 1$. The roots of the characteristic equation $r^2 - r - 1 = 0$ are $r_1 = (1 + \sqrt{5})/2$ and $r_2 = (1 - \sqrt{5})/2$. Therefore, from Theorem 1 it follows that the Fibonacci numbers are given by

$$f_n = \alpha_1 \left(\frac{1 + \sqrt{5}}{2}\right)^n + \alpha_2 \left(\frac{1 - \sqrt{5}}{2}\right)^n,$$

for some constants α_1 and α_2. The initial conditions $f_0 = 0$ and $f_1 = 1$ can be used to find these constants. We have

$$f_0 = \alpha_1 + \alpha_2 = 0,$$
$$f_1 = \alpha_1 \left(\frac{1+\sqrt{5}}{2}\right) + \alpha_2 \left(\frac{1-\sqrt{5}}{2}\right) = 1.$$

The solution to these simultaneous equations for α_1 and α_2 is

$$\alpha_1 = 1/\sqrt{5}, \qquad \alpha_2 = -1/\sqrt{5}.$$

Consequently, the Fibonacci numbers are given by

$$f_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n.$$

◀

Theorem 1 does not apply when there is one characteristic root of multiplicity two. If this happens, then $a_n = nr_0^n$ is another solution of the recurrence relation when r_0 is a root of multiplicity two of the characteristic equation. Theorem 2 shows how to handle this case.

THEOREM 2 Let c_1 and c_2 be real numbers with $c_2 \neq 0$. Suppose that $r^2 - c_1 r - c_2 = 0$ has only one root r_0. A sequence $\{a_n\}$ is a solution of the recurrence relation $a_n = c_1 a_{n-1} + c_2 a_{n-2}$ if and only if $a_n = \alpha_1 r_0^n + \alpha_2 n r_0^n$, for $n = 0, 1, 2, \ldots$, where α_1 and α_2 are constants.

The proof of Theorem 2 is left as an exercise at the end of the section. Example 5 illustrates the use of this theorem.

EXAMPLE 5 What is the solution of the recurrence relation

$$a_n = 6a_{n-1} - 9a_{n-2}$$

with initial conditions $a_0 = 1$ and $a_1 = 6$?

Solution: The only root of $r^2 - 6r + 9 = 0$ is $r = 3$. Hence, the solution to this recurrence relation is

$$a_n = \alpha_1 3^n + \alpha_2 n 3^n$$

for some constants α_1 and α_2. Using the initial conditions, it follows that

$$a_0 = 1 = \alpha_1,$$
$$a_1 = 6 = \alpha_1 \cdot 3 + \alpha_2 \cdot 3.$$

Solving these two equations shows that $\alpha_1 = 1$ and $\alpha_2 = 1$. Consequently, the solution to this recurrence relation and the initial conditions is

$$a_n = 3^n + n3^n.$$

◀

We will now state the general result about the solution of linear homogeneous recurrence relations with constant coefficients, where the degree may be greater than two, under the

assumption that the characteristic equation has distinct roots. The proof of this result will be left as an exercise for the reader.

THEOREM 3 Let $c_1, c_2, \ldots, c_k$ be real numbers. Suppose that the characteristic equation

$$r^k - c_1 r^{k-1} - \cdots - c_k = 0$$

has k distinct roots $r_1, r_2, \ldots, r_k$. Then a sequence $\{a_n\}$ is a solution of the recurrence relation

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_k a_{n-k}$$

if and only if

$$a_n = \alpha_1 r_1^n + \alpha_2 r_2^n + \cdots + \alpha_k r_k^n$$

for $n = 0, 1, 2, \ldots$, where $\alpha_1, \alpha_2, \ldots, \alpha_k$ are constants.

We illustrate the use of the theorem with Example 6.

EXAMPLE 6 Find the solution to the recurrence relation

$$a_n = 6a_{n-1} - 11a_{n-2} + 6a_{n-3}$$

with the initial conditions $a_0 = 2$, $a_1 = 5$, and $a_2 = 15$.

Solution: The characteristic polynomial of this recurrence relation is

$$r^3 - 6r^2 + 11r - 6.$$

The characteristic roots are $r = 1$, $r = 2$, and $r = 3$, because $r^3 - 6r^2 + 11r - 6 = (r - 1)(r - 2)(r - 3)$. Hence, the solutions to this recurrence relation are of the form

$$a_n = \alpha_1 \cdot 1^n + \alpha_2 \cdot 2^n + \alpha_3 \cdot 3^n.$$

To find the constants α_1, α_2, and α_3, use the initial conditions. This gives

$$\begin{aligned} a_0 &= 2 = \alpha_1 + \alpha_2 + \alpha_3, \\ a_1 &= 5 = \alpha_1 + \alpha_2 \cdot 2 + \alpha_3 \cdot 3, \\ a_2 &= 15 = \alpha_1 + \alpha_2 \cdot 4 + \alpha_3 \cdot 9. \end{aligned}$$

When these three simultaneous equations are solved for α_1, α_2, and α_3, we find that $\alpha_1 = 1$, $\alpha_2 = -1$, and $\alpha_3 = 2$. Hence, the unique solution to this recurrence relation and the given initial conditions is the sequence $\{a_n\}$ with

$$a_n = 1 - 2^n + 2 \cdot 3^n.$$ ◀

We now state the most general result about linear homogeneous recurrence relations with constant coefficients, allowing the characteristic equation to have multiple roots. The key point is that for each root r of the characteristic equation, the general solution has a summand of the

form $P(n)r^n$, where $P(n)$ is a polynomial of degree $m-1$, with m the multiplicity of this root. We leave the proof of this result as a challenging exercise for the reader.

THEOREM 4 Let $c_1, c_2, \ldots, c_k$ be real numbers. Suppose that the characteristic equation

$$r^k - c_1 r^{k-1} - \cdots - c_k = 0$$

has t distinct roots $r_1, r_2, \ldots, r_t$ with multiplicities $m_1, m_2, \ldots, m_t$, respectively, so that $m_i \geq 1$ for $i = 1, 2, \ldots, t$ and $m_1 + m_2 + \cdots + m_t = k$. Then a sequence $\{a_n\}$ is a solution of the recurrence relation

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_k a_{n-k}$$

if and only if

$$\begin{aligned} a_n = {} & (\alpha_{1,0} + \alpha_{1,1} n + \cdots + \alpha_{1,m_1-1} n^{m_1-1}) r_1^n \\ & + (\alpha_{2,0} + \alpha_{2,1} n + \cdots + \alpha_{2,m_2-1} n^{m_2-1}) r_2^n \\ & + \cdots + (\alpha_{t,0} + \alpha_{t,1} n + \cdots + \alpha_{t,m_t-1} n^{m_t-1}) r_t^n \end{aligned}$$

for $n = 0, 1, 2, \ldots$, where $\alpha_{i,j}$ are constants for $1 \leq i \leq t$ and $0 \leq j \leq m_i - 1$.

Example 7 illustrates how Theorem 4 is used to find the general form of a solution of a linear homogeneous recurrence relation when the characteristic equation has several repeated roots.

EXAMPLE 7 Suppose that the roots of the characteristic equation of a linear homogeneous recurrence relation are 2, 2, 2, 5, 5, and 9 (that is, there are three roots, the root 2 with multiplicity three, the root 5 with multiplicity two, and the root 9 with multiplicity one). What is the form of the general solution?

Solution: By Theorem 4, the general form of the solution is

$$(\alpha_{1,0} + \alpha_{1,1} n + \alpha_{1,2} n^2) 2^n + (\alpha_{2,0} + \alpha_{2,1} n) 5^n + \alpha_{3,0} 9^n. \quad \blacktriangleleft$$

We now illustrate the use of Theorem 4 to solve a linear homogeneous recurrence relation with constant coefficients when the characteristic equation has a root of multiplicity three.

EXAMPLE 8 Find the solution to the recurrence relation

$$a_n = -3a_{n-1} - 3a_{n-2} - a_{n-3}$$

with initial conditions $a_0 = 1$, $a_1 = -2$, and $a_2 = -1$.

Solution: The characteristic equation of this recurrence relation is

$$r^3 + 3r^2 + 3r + 1 = 0.$$

Because $r^3 + 3r^2 + 3r + 1 = (r + 1)^3$, there is a single root $r = -1$ of multiplicity three of the characteristic equation. By Theorem 4 the solutions of this recurrence relation are of the form

$$a_n = \alpha_{1,0}(-1)^n + \alpha_{1,1}n(-1)^n + \alpha_{1,2}n^2(-1)^n.$$

To find the constants $\alpha_{1,0}$, $\alpha_{1,1}$, and $\alpha_{1,2}$, use the initial conditions. This gives

$$\begin{aligned} a_0 &= 1 = \alpha_{1,0}, \\ a_1 &= -2 = -\alpha_{1,0} - \alpha_{1,1} - \alpha_{1,2}, \\ a_2 &= -1 = \alpha_{1,0} + 2\alpha_{1,1} + 4\alpha_{1,2}. \end{aligned}$$

The simultaneous solution of these three equations is $\alpha_{1,0} = 1, \alpha_{1,1} = 3$, and $\alpha_{1,2} = -2$. Hence, the unique solution to this recurrence relation and the given initial conditions is the sequence $\{a_n\}$ with

$$a_n = (1 + 3n - 2n^2)(-1)^n.$$ ◀

Linear Nonhomogeneous Recurrence Relations with Constant Coefficients

We have seen how to solve linear homogeneous recurrence relations with constant coefficients. Is there a relatively simple technique for solving a linear, but not homogeneous, recurrence relation with constant coefficients, such as $a_n = 3a_{n-1} + 2n$? We will see that the answer is yes for certain families of such recurrence relations.

The recurrence relation $a_n = 3a_{n-1} + 2n$ is an example of a **linear nonhomogeneous recurrence relation with constant coefficients,** that is, a recurrence relation of the form

$$a_n = c_1a_{n-1} + c_2a_{n-2} + \cdots + c_ka_{n-k} + F(n),$$

where $c_1, c_2, \ldots, c_k$ are real numbers and $F(n)$ is a function not identically zero depending only on n. The recurrence relation

$$a_n = c_1a_{n-1} + c_2a_{n-2} + \cdots + c_ka_{n-k}$$

is called the **associated homogeneous recurrence relation.** It plays an important role in the solution of the nonhomogeneous recurrence relation.

EXAMPLE 9 Each of the recurrence relations $a_n = a_{n-1} + 2^n$, $a_n = a_{n-1} + a_{n-2} + n^2 + n + 1$, $a_n = 3a_{n-1} + n3^n$, and $a_n = a_{n-1} + a_{n-2} + a_{n-3} + n!$ is a linear nonhomogeneous recurrence relation with constant coefficients. The associated linear homogeneous recurrence relations are $a_n = a_{n-1}$, $a_n = a_{n-1} + a_{n-2}$, $a_n = 3a_{n-1}$, and $a_n = a_{n-1} + a_{n-2} + a_{n-3}$, respectively. ◀

The key fact about linear nonhomogeneous recurrence relations with constant coefficients is that every solution is the sum of a particular solution and a solution of the associated linear homogeneous recurrence relation, as Theorem 5 shows.

THEOREM 5 If $\{a_n^{(p)}\}$ is a particular solution of the nonhomogeneous linear recurrence relation with constant coefficients

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_k a_{n-k} + F(n),$$

then every solution is of the form $\{a_n^{(p)} + a_n^{(h)}\}$, where $\{a_n^{(h)}\}$ is a solution of the associated homogeneous recurrence relation

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_k a_{n-k}.$$

Proof: Because $\{a_n^{(p)}\}$ is a particular solution of the nonhomogeneous recurrence relation, we know that

$$a_n^{(p)} = c_1 a_{n-1}^{(p)} + c_2 a_{n-2}^{(p)} + \cdots + c_k a_{n-k}^{(p)} + F(n).$$

Now suppose that $\{b_n\}$ is a second solution of the nonhomogeneous recurrence relation, so that

$$b_n = c_1 b_{n-1} + c_2 b_{n-2} + \cdots + c_k b_{n-k} + F(n).$$

Subtracting the first of these two equations from the second shows that

$$b_n - a_n^{(p)} = c_1(b_{n-1} - a_{n-1}^{(p)}) + c_2(b_{n-2} - a_{n-2}^{(p)}) + \cdots + c_k(b_{n-k} - a_{n-k}^{(p)}).$$

It follows that $\{b_n - a_n^{(p)}\}$ is a solution of the associated homogeneous linear recurrence, say, $\{a_n^{(h)}\}$. Consequently, $b_n = a_n^{(p)} + a_n^{(h)}$ for all n. ◁

By Theorem 5, we see that the key to solving nonhomogeneous recurrence relations with constant coefficients is finding a particular solution. Then every solution is a sum of this solution and a solution of the associated homogeneous recurrence relation. Although there is no general method for finding such a solution that works for every function $F(n)$, there are techniques that work for certain types of functions $F(n)$, such as polynomials and powers of constants. This is illustrated in Examples 10 and 11.

EXAMPLE 10 Find all solutions of the recurrence relation $a_n = 3a_{n-1} + 2n$. What is the solution with $a_1 = 3$?

Solution: To solve this linear nonhomogeneous recurrence relation with constant coefficients, we need to solve its associated linear homogeneous equation and to find a particular solution for the given nonhomogeneous equation. The associated linear homogeneous equation is $a_n = 3a_{n-1}$. Its solutions are $a_n^{(h)} = \alpha 3^n$, where α is a constant.

We now find a particular solution. Because $F(n) = 2n$ is a polynomial in n of degree one, a reasonable trial solution is a linear function in n, say, $p_n = cn + d$, where c and d are constants. To determine whether there are any solutions of this form, suppose that $p_n = cn + d$ is such a solution. Then the equation $a_n = 3a_{n-1} + 2n$ becomes $cn + d = 3(c(n-1) + d) + 2n$. Simplifying and combining like terms gives $(2 + 2c)n + (2d - 3c) = 0$. It follows that $cn + d$ is a solution if and only if $2 + 2c = 0$ and $2d - 3c = 0$. This shows that $cn + d$ is a solution if and only if $c = -1$ and $d = -3/2$. Consequently, $a_n^{(p)} = -n - 3/2$ is a particular solution.

By Theorem 5 all solutions are of the form

$$a_n = a_n^{(p)} + a_n^{(h)} = -n - \frac{3}{2} + \alpha \cdot 3^n,$$

where α is a constant.

To find the solution with $a_1 = 3$, let $n = 1$ in the formula we obtained for the general solution. We find that $3 = -1 - 3/2 + 3\alpha$, which implies that $\alpha = 11/6$. The solution we seek is $a_n = -n - 3/2 + (11/6)3^n$. ◀

EXAMPLE 11 Find all solutions of the recurrence relation

Extra Examples

$$a_n = 5a_{n-1} - 6a_{n-2} + 7^n.$$

Solution: This is a linear nonhomogeneous recurrence relation. The solutions of its associated homogeneous recurrence relation

$$a_n = 5a_{n-1} - 6a_{n-2}$$

are $a_n^{(h)} = \alpha_1 \cdot 3^n + \alpha_2 \cdot 2^n$, where α_1 and α_2 are constants. Because $F(n) = 7^n$, a reasonable trial solution is $a_n^{(p)} = C \cdot 7^n$, where C is a constant. Substituting the terms of this sequence into the recurrence relation implies that $C \cdot 7^n = 5C \cdot 7^{n-1} - 6C \cdot 7^{n-2} + 7^n$. Factoring out 7^{n-2}, this equation becomes $49C = 35C - 6C + 49$, which implies that $20C = 49$, or that $C = 49/20$. Hence, $a_n^{(p)} = (49/20)7^n$ is a particular solution. By Theorem 5, all solutions are of the form

$$a_n = \alpha_1 \cdot 3^n + \alpha_2 \cdot 2^n + (49/20)7^n.$$ ◀

In Examples 10 and 11, we made an educated guess that there are solutions of a particular form. In both cases we were able to find particular solutions. This was not an accident. Whenever $F(n)$ is the product of a polynomial in n and the nth power of a constant, we know exactly what form a particular solution has, as stated in Theorem 6. We leave the proof of Theorem 6 as a challenging exercise for the reader.

THEOREM 6 Suppose that $\{a_n\}$ satisfies the linear nonhomogeneous recurrence relation

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_k a_{n-k} + F(n),$$

where $c_1, c_2, \ldots, c_k$ are real numbers, and

$$F(n) = (b_t n^t + b_{t-1} n^{t-1} + \cdots + b_1 n + b_0)s^n,$$

where $b_0, b_1, \ldots, b_t$ and s are real numbers. When s is not a root of the characteristic equation of the associated linear homogeneous recurrence relation, there is a particular solution of the form

$$(p_t n^t + p_{t-1} n^{t-1} + \cdots + p_1 n + p_0)s^n.$$

When s is a root of this characteristic equation and its multiplicity is m, there is a particular solution of the form

$$n^m(p_t n^t + p_{t-1} n^{t-1} + \cdots + p_1 n + p_0)s^n.$$

Note that in the case when s is a root of multiplicity m of the characteristic equation of the associated linear homogeneous recurrence relation, the factor n^m ensures that the proposed particular solution will not already be a solution of the associated linear homogeneous recurrence

relation. We next provide Example 12 to illustrate the form of a particular solution provided by Theorem 6.

EXAMPLE 12 What form does a particular solution of the linear nonhomogeneous recurrence relation $a_n = 6a_{n-1} - 9a_{n-2} + F(n)$ have when $F(n) = 3^n$, $F(n) = n3^n$, $F(n) = n^2 2^n$, and $F(n) = (n^2 + 1)3^n$?

Solution: The associated linear homogeneous recurrence relation is $a_n = 6a_{n-1} - 9a_{n-2}$. Its characteristic equation, $r^2 - 6r + 9 = (r - 3)^2 = 0$, has a single root, 3, of multiplicity two. To apply Theorem 6, with $F(n)$ of the form $P(n)s^n$, where $P(n)$ is a polynomial and s is a constant, we need to ask whether s is a root of this characteristic equation.

Because $s = 3$ is a root with multiplicity $m = 2$ but $s = 2$ is not a root, Theorem 6 tells us that a particular solution has the form $p_0 n^2 3^n$ if $F(n) = 3^n$, the form $n^2(p_1 n + p_0)3^n$ if $F(n) = n3^n$, the form $(p_2 n^2 + p_1 n + p_0)2^n$ if $F(n) = n^2 2^n$, and the form $n^2(p_2 n^2 + p_1 n + p_0)3^n$ if $F(n) = (n^2 + 1)3^n$. ◀

Care must be taken when $s = 1$ when solving recurrence relations of the type covered by Theorem 6. In particular, to apply this theorem with $F(n) = b_t n_t + b_{t-1} n_{t-1} + \cdots + b_1 n + b_0$, the parameter s takes the value $s = 1$ (even though the term 1^n does not explicitly appear). By the theorem, the form of the solution then depends on whether 1 is a root of the characteristic equation of the associated linear homogeneous recurrence relation. This is illustrated in Example 13, which shows how Theorem 6 can be used to find a formula for the sum of the first n positive integers.

EXAMPLE 13 Let a_n be the sum of the first n positive integers, so that

$$a_n = \sum_{k=1}^{n} k.$$

Note that a_n satisfies the linear nonhomogeneous recurrence relation

$$a_n = a_{n-1} + n.$$

(To obtain a_n, the sum of the first n positive integers, from a_{n-1}, the sum of the first $n - 1$ positive integers, we add n.) Note that the initial condition is $a_1 = 1$.

The associated linear homogeneous recurrence relation for a_n is

$$a_n = a_{n-1}.$$

The solutions of this homogeneous recurrence relation are given by $a_n^{(h)} = c(1)^n = c$, where c is a constant. To find all solutions of $a_n = a_{n-1} + n$, we need find only a single particular solution. By Theorem 6, because $F(n) = n = n \cdot (1)^n$ and $s = 1$ is a root of degree one of the characteristic equation of the associated linear homogeneous recurrence relations, there is a particular solution of the form $n(p_1 n + p_0) = p_1 n^2 + p_0 n$.

Inserting this into the recurrence relation gives $p_1 n^2 + p_0 n = p_1(n - 1)^2 + p_0(n - 1) + n$. Simplifying, we see that $n(2p_1 - 1) + (p_0 - p_1) = 0$, which means that $2p_1 - 1 = 0$ and $p_0 - p_1 = 0$, so $p_0 = p_1 = 1/2$. Hence,

$$a_n^{(p)} = \frac{n^2}{2} + \frac{n}{2} = \frac{n(n+1)}{2}$$

is a particular solution. Hence, all solutions of the original recurrence relation $a_n = a_{n-1} + n$ are given by $a_n = a_n^{(h)} + a_n^{(p)} = c + n(n+1)/2$. Because $a_1 = 1$, we have $1 = a_1 = c + 1 \cdot 2/2 = c + 1$, so $c = 0$. It follows that $a_n = n(n+1)/2$. (This is the same formula given in Table 2 in Section 2.4 and derived previously.) ◀

Exercises

1. Determine which of these are linear homogeneous recurrence relations with constant coefficients. Also, find the degree of those that are.
   a) $a_n = 3a_{n-1} + 4a_{n-2} + 5a_{n-3}$
   b) $a_n = 2na_{n-1} + a_{n-2}$ c) $a_n = a_{n-1} + a_{n-4}$
   d) $a_n = a_{n-1} + 2$ e) $a_n = a_{n-1}^2 + a_{n-2}$
   f) $a_n = a_{n-2}$ g) $a_n = a_{n-1} + n$
2. Determine which of these are linear homogeneous recurrence relations with constant coefficients. Also, find the degree of those that are.
   a) $a_n = 3a_{n-2}$ b) $a_n = 3$
   c) $a_n = a_{n-1}^2$ d) $a_n = a_{n-1} + 2a_{n-3}$
   e) $a_n = a_{n-1}/n$
   f) $a_n = a_{n-1} + a_{n-2} + n + 3$
   g) $a_n = 4a_{n-2} + 5a_{n-4} + 9a_{n-7}$
3. Solve these recurrence relations together with the initial conditions given.
   a) $a_n = 2a_{n-1}$ for $n \geq 1$, $a_0 = 3$
   b) $a_n = a_{n-1}$ for $n \geq 1$, $a_0 = 2$
   c) $a_n = 5a_{n-1} - 6a_{n-2}$ for $n \geq 2$, $a_0 = 1$, $a_1 = 0$
   d) $a_n = 4a_{n-1} - 4a_{n-2}$ for $n \geq 2$, $a_0 = 6$, $a_1 = 8$
   e) $a_n = -4a_{n-1} - 4a_{n-2}$ for $n \geq 2$, $a_0 = 0$, $a_1 = 1$
   f) $a_n = 4a_{n-2}$ for $n \geq 2$, $a_0 = 0$, $a_1 = 4$
   g) $a_n = a_{n-2}/4$ for $n \geq 2$, $a_0 = 1$, $a_1 = 0$
4. Solve these recurrence relations together with the initial conditions given.
   a) $a_n = a_{n-1} + 6a_{n-2}$ for $n \geq 2$, $a_0 = 3$, $a_1 = 6$
   b) $a_n = 7a_{n-1} - 10a_{n-2}$ for $n \geq 2$, $a_0 = 2$, $a_1 = 1$
   c) $a_n = 6a_{n-1} - 8a_{n-2}$ for $n \geq 2$, $a_0 = 4$, $a_1 = 10$
   d) $a_n = 2a_{n-1} - a_{n-2}$ for $n \geq 2$, $a_0 = 4$, $a_1 = 1$
   e) $a_n = a_{n-2}$ for $n \geq 2$, $a_0 = 5$, $a_1 = -1$
   f) $a_n = -6a_{n-1} - 9a_{n-2}$ for $n \geq 2$, $a_0 = 3$, $a_1 = -3$
   g) $a_{n+2} = -4a_{n+1} + 5a_n$ for $n \geq 0$, $a_0 = 2$, $a_1 = 8$
5. How many different messages can be transmitted in n microseconds using the two signals described in Exercise 35 in Section 7.1?
6. How many different messages can be transmitted in n microseconds using three different signals if one signal requires 1 microsecond for transmittal, the other two signals require 2 microseconds each for transmittal, and a signal in a message is followed immediately by the next signal?
7. In how many ways can a $2 \times n$ rectangular checkerboard be tiled using 1×2 and 2×2 pieces?
8. A model for the number of lobsters caught per year is based on the assumption that the number of lobsters caught in a year is the average of the number caught in the two previous years.
   a) Find a recurrence relation for $\{L_n\}$, where L_n is the number of lobsters caught in year n, under the assumption for this model.
   b) Find L_n if 100,000 lobsters were caught in year 1 and 300,000 were caught in year 2.
9. A deposit of \$100,000 is made to an investment fund at the beginning of a year. On the last day of each year two dividends are awarded. The first dividend is 20% of the amount in the account during that year. The second dividend is 45% of the amount in the account in the previous year.
   a) Find a recurrence relation for $\{P_n\}$, where P_n is the amount in the account at the end of n years if no money is ever withdrawn.
   b) How much is in the account after n years if no money has been withdrawn?

*10. Prove Theorem 2.

11. The **Lucas numbers** satisfy the recurrence relation

$$L_n = L_{n-1} + L_{n-2},$$

and the initial conditions $L_0 = 2$ and $L_1 = 1$.
   a) Show that $L_n = f_{n-1} + f_{n+1}$ for $n = 2, 3, \ldots$, where f_n is the nth Fibonacci number.
   b) Find an explicit formula for the Lucas numbers.
12. Find the solution to $a_n = 2a_{n-1} + a_{n-2} - 2a_{n-3}$ for $n = 3, 4, 5, \ldots$, with $a_0 = 3$, $a_1 = 6$, and $a_2 = 0$.
13. Find the solution to $a_n = 7a_{n-2} + 6a_{n-3}$ with $a_0 = 9$, $a_1 = 10$, and $a_2 = 32$.
14. Find the solution to $a_n = 5a_{n-2} - 4a_{n-4}$ with $a_0 = 3$, $a_1 = 2$, $a_2 = 6$, and $a_3 = 8$.
15. Find the solution to $a_n = 2a_{n-1} + 5a_{n-2} - 6a_{n-3}$ with $a_0 = 7$, $a_1 = -4$, and $a_2 = 8$.

*16. Prove Theorem 3.

17. Prove this identity relating the Fibonacci numbers and the binomial coefficients:

$$f_{n+1} = C(n, 0) + C(n-1, 1) + \cdots + C(n-k, k),$$

where n is a positive integer and $k = \lfloor n/2 \rfloor$. [*Hint:* Let $a_n = C(n, 0) + C(n-1, 1) + \cdots + C(n-k, k)$. Show that the sequence $\{a_n\}$ satisfies the same recurrence relation and initial conditions satisfied by the sequence of Fibonacci numbers.]

18. Solve the recurrence relation $a_n = 6a_{n-1} - 12a_{n-2} + 8a_{n-3}$ with $a_0 = -5$, $a_1 = 4$, and $a_2 = 88$.

19. Solve the recurrence relation $a_n = -3a_{n-1} - 3a_{n-2} - a_{n-3}$ with $a_0 = 5$, $a_1 = -9$, and $a_2 = 15$.

20. Find the general form of the solutions of the recurrence relation $a_n = 8a_{n-2} - 16a_{n-4}$.

21. What is the general form of the solutions of a linear homogeneous recurrence relation if its characteristic equation has roots 1, 1, 1, 1, −2, −2, −2, 3, 3, −4?

22. What is the general form of the solutions of a linear homogeneous recurrence relation if its characteristic equation has the roots −1, −1, −1, 2, 2, 5, 5, 7?

23. Consider the nonhomogeneous linear recurrence relation $a_n = 3a_{n-1} + 2^n$.

a) Show that $a_n = -2^{n+1}$ is a solution of this recurrence relation.

b) Use Theorem 5 to find all solutions of this recurrence relation.

c) Find the solution with $a_0 = 1$.

24. Consider the nonhomogeneous linear recurrence relation $a_n = 2a_{n-1} + 2^n$.

a) Show that $a_n = n2^n$ is a solution of this recurrence relation.

b) Use Theorem 5 to find all solutions of this recurrence relation.

c) Find the solution with $a_0 = 2$.

25. a) Determine values of the constants A and B such that $a_n = An + B$ is a solution of recurrence relation $a_n = 2a_{n-1} + n + 5$.

b) Use Theorem 5 to find all solutions of this recurrence relation.

c) Find the solution of this recurrence relation with $a_0 = 4$.

26. What is the general form of the particular solution guaranteed to exist by Theorem 6 of the linear nonhomogeneous recurrence relation $a_n = 6a_{n-1} - 12a_{n-2} + 8a_{n-3} + F(n)$ if

a) $F(n) = n^2$? **b)** $F(n) = 2^n$?
c) $F(n) = n2^n$? **d)** $F(n) = (-2)^n$?
e) $F(n) = n^2 2^n$? **f)** $F(n) = n^3(-2)^n$?
g) $F(n) = 3$?

27. What is the general form of the particular solution guaranteed to exist by Theorem 6 of the linear nonhomogeneous recurrence relation $a_n = 8a_{n-2} - 16a_{n-4} + F(n)$ if

a) $F(n) = n^3$? **b)** $F(n) = (-2)^n$?
c) $F(n) = n2^n$? **d)** $F(n) = n^2 4^n$?
e) $F(n) = (n^2 - 2)(-2)^n$? **f)** $F(n) = n^4 2^n$?
g) $F(n) = 2$?

28. a) Find all solutions of the recurrence relation $a_n = 2a_{n-1} + 2n^2$.

b) Find the solution of the recurrence relation in part (a) with initial condition $a_1 = 4$.

29. a) Find all solutions of the recurrence relation $a_n = 2a_{n-1} + 3^n$.

b) Find the solution of the recurrence relation in part (a) with initial condition $a_1 = 5$.

30. a) Find all solutions of the recurrence relation $a_n = -5a_{n-1} - 6a_{n-2} + 42 \cdot 4^n$.

b) Find the solution of this recurrence relation with $a_1 = 56$ and $a_2 = 278$.

31. Find all solutions of the recurrence relation $a_n = 5a_{n-1} - 6a_{n-2} + 2^n + 3n$. [*Hint:* Look for a particular solution of the form $qn2^n + p_1 n + p_2$, where q, p_1, and p_2 are constants.]

32. Find the solution of the recurrence relation $a_n = 2a_{n-1} + 3 \cdot 2^n$.

33. Find all solutions of the recurrence relation $a_n = 4a_{n-1} - 4a_{n-2} + (n+1)2^n$.

34. Find all solutions of the recurrence relation $a_n = 7a_{n-1} - 16a_{n-2} + 12a_{n-3} + n4^n$ with $a_0 = -2$, $a_1 = 0$, and $a_2 = 5$.

35. Find the solution of the recurrence relation $a_n = 4a_{n-1} - 3a_{n-2} + 2^n + n + 3$ with $a_0 = 1$ and $a_1 = 4$.

36. Let a_n be the sum of the first n perfect squares, that is, $a_n = \sum_{k=1}^{n} k^2$. Show that the sequence $\{a_n\}$ satisfies the linear nonhomogeneous recurrence relation $a_n = a_{n-1} + n^2$ and the initial condition $a_1 = 1$. Use Theorem 6 to determine a formula for a_n by solving this recurrence relation.

37. Let a_n be the sum of the first n triangular numbers, that is, $a_n = \sum_{k=1}^{n} t_k$, where $t_k = k(k+1)/2$. Show that $\{a_n\}$ satisfies the linear nonhomogeneous recurrence relation $a_n = a_{n-1} + n(n+1)/2$ and the initial condition $a_1 = 1$. Use Theorem 6 to determine a formula for a_n by solving this recurrence relation.

38. a) Find the characteristic roots of the linear homogeneous recurrence relation $a_n = 2a_{n-1} - 2a_{n-2}$. (*Note:* These are complex numbers.)

b) Find the solution of the recurrence relation in part (a) with $a_0 = 1$ and $a_1 = 2$.

***39. a)** Find the characteristic roots of the linear homogeneous recurrence relation $a_n = a_{n-4}$. (*Note:* These include complex numbers.)

b) Find the solution of the recurrence relation in part (a) with $a_0 = 1$, $a_1 = 0$, $a_2 = -1$, and $a_3 = 1$.

***40.** Solve the simultaneous recurrence relations

$$a_n = 3a_{n-1} + 2b_{n-1}$$
$$b_n = a_{n-1} + 2b_{n-1}$$

with $a_0 = 1$ and $b_0 = 2$.

***41. a)** Use the formula found in Example 4 for f_n, the nth Fibonacci number, to show that f_n is the integer closest to

$$\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n.$$

b) Determine for which n f_n is greater than

$$\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n$$

and for which n f_n is less than

$$\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n.$$

42. Show that if $a_n = a_{n-1} + a_{n-2}$, $a_0 = s$ and $a_1 = t$, where s and t are constants, then $a_n = sf_{n-1} + tf_n$ for all positive integers n.

43. Express the solution of the linear nonhomogenous recurrence relation $a_n = a_{n-1} + a_{n-2} + 1$ for $n \geq 2$ where $a_0 = 0$ and $a_1 = 1$ in terms of the Fibonacci numbers. [*Hint:* Let $b_n = a_n + 1$ and apply Exercise 42 to the sequence b_n.]

***44.** (*Linear algebra required*) Let $\mathbf{A}_n$ be the $n \times n$ matrix with 2s on its main diagonal, 1s in all positions next to a diagonal element, and 0s everywhere else. Find a recurrence relation for d_n, the determinant of $\mathbf{A}_n$. Solve this recurrence relation to find a formula for d_n.

45. Suppose that each pair of a genetically engineered species of rabbits left on an island produces two new pairs of rabbits at the age of 1 month and six new pairs of rabbits at the age of 2 months and every month afterward. None of the rabbits ever die or leave the island.

a) Find a recurrence relation for the number of pairs of rabbits on the island n months after one newborn pair is left on the island.

b) By solving the recurrence relation in (a) determine the number of pairs of rabbits on the island n months after one pair is left on the island.

46. Suppose that there are two goats on an island initially. The number of goats on the island doubles every year by natural reproduction, and some goats are either added or removed each year.

a) Construct a recurrence relation for the number of goats on the island at the start of the nth year, assuming that during each year an extra 100 goats are put on the island.

b) Solve the recurrence relation from part (a) to find the number of goats on the island at the start of the nth year.

c) Construct a recurrence relation for the number of goats on the island at the start of the nth year, assuming that n goats are removed during the nth year for each $n \geq 3$.

d) Solve the recurrence relation in part (c) for the number of goats on the island at the start of the nth year.

47. A new employee at an exciting new software company starts with a salary of \$50,000 and is promised that at the end of each year her salary will be double her salary of the previous year, with an extra increment of \$10,000 for each year she has been with the company.

a) Construct a recurrence relation for her salary for her nth year of employment.

b) Solve this recurrence relation to find her salary for her nth year of employment.

Some linear recurrence relations that do not have constant coefficients can be systematically solved. This is the case for recurrence relations of the form $f(n)a_n = g(n)a_{n-1} + h(n)$. Exercises 48–50 illustrate this.

***48. a)** Show that the recurrence relation

$$f(n)a_n = g(n)a_{n-1} + h(n),$$

for $n \geq 1$, and with $a_0 = C$, can be reduced to a recurrence relation of the form

$$b_n = b_{n-1} + Q(n)h(n),$$

where $b_n = g(n+1)Q(n+1)a_n$, with

$$Q(n) = (f(1)f(2)\cdots f(n-1))/(g(1)g(2)\cdots g(n)).$$

b) Use part (a) to solve the original recurrence relation to obtain

$$a_n = \frac{C + \sum_{i=1}^{n} Q(i)h(i)}{g(n+1)Q(n+1)}.$$

***49.** Use Exercise 48 to solve the recurrence relation $(n+1)a_n = (n+3)a_{n-1} + n$, for $n \geq 1$, with $a_0 = 1$.

50. It can be shown that the average number of comparisons made by the quick sort algorithm (described in preamble to Exercise 50 in Section 4.4), when sorting n elements in random order, satisfies the recurrence relation

$$C_n = n + 1 + \frac{2}{n}\sum_{k=0}^{n-1} C_k$$

for $n = 1, 2, \ldots$, with initial condition $C_0 = 0$.

a) Show that $\{C_n\}$ also satisfies the recurrence relation $nC_n = (n+1)C_{n-1} + 2n$ for $n = 1, 2, \ldots$.

b) Use Exercise 48 to solve the recurrence relation in part (a) to find an explicit formula for C_n.

****51.** Prove Theorem 4.

****52.** Prove Theorem 6.

53. Solve the recurrence relation $T(n) = nT^2(n/2)$ with initial condition $T(1) = 6$. [*Hint:* Let $n = 2^k$ and then make the substitution $a_k = \log T(2^k)$ to obtain a linear nonhomogeneous recurrence relation.]

7.3 Divide-and-Conquer Algorithms and Recurrence Relations

Introduction

Links

Many recursive algorithms take a problem with a given input and divide it into one or more smaller problems. This reduction is successively applied until the solutions of the smaller problems can be found quickly. For instance, we perform a binary search by reducing the search for an element in a list to the search for this element in a list half as long. We successively apply this reduction until one element is left. When we sort a list of integers using the merge sort, we split the list into two halves of equal size and sort each half separately. We then merge the two sorted halves. Another example of this type of recursive algorithm is a procedure for multiplying integers that reduces the problem of the multiplication of two integers to three multiplications of pairs of integers with half as many bits. This reduction is successively applied until integers with one bit are obtained. These procedures are called **divide-and-conquer algorithms** because they *divide* a problem into one or more instances of the same problem of smaller size and they *conquer* the problem by using the solutions of the smaller problems to find a solution of the original problem, perhaps with some additional work.

In this section we will show how recurrence relations can be used to analyze the computational complexity of divide-and-conquer algorithms. We will use these recurrence relations to estimate the number of operations used by many different divide-and-conquer algorithms, including several that we introduce in this section.

Divide-and-Conquer Recurrence Relations

Suppose that a recursive algorithm divides a problem of size n into a subproblems, where each subproblem is of size n/b (for simplicity, assume that n is a multiple of b; in reality, the smaller problems are often of size equal to the nearest integers either less than or equal to, or greater than or equal to, n/b). Also, suppose that a total of $g(n)$ extra operations are required in the conquer step of the algorithm to combine the solutions of the subproblems into a solution of the original problem. Then, if $f(n)$ represents the number of operations required to solve the problem of size n, it follows that f satisfies the recurrence relation

$$f(n) = af(n/b) + g(n).$$

This is called a **divide-and-conquer recurrence relation.**

We will first set up the divide-and-conquer recurrence relations that can be used to study the complexity of some important algorithms. Then we will show how to use these divide-and-conquer recurrence relations to estimate the complexity of these algorithms.

EXAMPLE 1

Extra Examples

Binary Search We introduced a binary search algorithm in Section 3.1. This binary search algorithm reduces the search for an element in a search sequence of size n to the binary search for this element in a search sequence of size $n/2$, when n is even. (Hence, the problem of size n has been reduced to *one* problem of size $n/2$.) Two comparisons are needed to implement this reduction (one to determine which half of the list to use and the other to determine whether any terms of the list remain). Hence, if $f(n)$ is the number of comparisons required to search for an element in a search sequence of size n, then

$$f(n) = f(n/2) + 2$$

when n is even. ◀

EXAMPLE 2 **Finding the Maximum and Minimum of a Sequence** Consider the following algorithm for locating the maximum and minimum elements of a sequence $a_1, a_2, \ldots, a_n$. If $n = 1$, then a_1 is the maximum and the minimum. If $n > 1$, split the sequence into two sequences, either where both have the same number of elements or where one of the sequences has one more element than the other. The problem is reduced to finding the maximum and minimum of each of the two smaller sequences. The solution to the original problem results from the comparison of the separate maxima and minima of the two smaller sequences to obtain the overall maximum and minimum.

Let $f(n)$ be the total number of comparisons needed to find the maximum and minimum elements of the sequence with n elements. We have shown that a problem of size n can be reduced into two problems of size $n/2$, when n is even, using two comparisons, one to compare the maxima of the two sequences and the other to compare the minima of the two sequences. This gives the recurrence relation

$$f(n) = 2f(n/2) + 2$$

when n is even. ◀

EXAMPLE 3 **Merge Sort** The merge sort algorithm (introduced in Section 4.4) splits a list to be sorted with n items, where n is even, into two lists with $n/2$ elements each, and uses fewer than n comparisons to merge the two sorted lists of $n/2$ items each into one sorted list. Consequently, the number of comparisons used by the merge sort to sort a list of n elements is less than $M(n)$, where the function $M(n)$ satisfies the divide-and-conquer recurrence relation

$$M(n) = 2M(n/2) + n.$$ ◀

EXAMPLE 4 **Fast Multiplication of Integers** Surprisingly, there are more efficient algorithms than the conventional algorithm (described in Section 3.6) for multiplying integers. One of these algorithms, which uses a divide-and-conquer technique, will be described here. This fast multiplication algorithm proceeds by splitting each of two $2n$-bit integers into two blocks, each with n bits. Then, the original multiplication is reduced from the multiplication of two $2n$-bit integers to three multiplications of n-bit integers, plus shifts and additions.

Links

Suppose that a and b are integers with binary expansions of length $2n$ (add initial bits of zero in these expansions if necessary to make them the same length). Let

$$a = (a_{2n-1}a_{2n-2}\cdots a_1a_0)_2 \qquad \text{and} \qquad b = (b_{2n-1}b_{2n-2}\cdots b_1b_0)_2.$$

Let

$$a = 2^n A_1 + A_0, \quad b = 2^n B_1 + B_0,$$

where

$$A_1 = (a_{2n-1}\cdots a_{n+1}a_n)_2, \qquad A_0 = (a_{n-1}\cdots a_1a_0)_2,$$
$$B_1 = (b_{2n-1}\cdots b_{n+1}b_n)_2, \qquad B_0 = (b_{n-1}\cdots b_1b_0)_2.$$

The algorithm for fast multiplication of integers is based on the fact that ab can be rewritten as

$$ab = (2^{2n} + 2^n)A_1B_1 + 2^n(A_1 - A_0)(B_0 - B_1) + (2^n + 1)A_0B_0.$$

The important fact about this identity is that it shows that the multiplication of two $2n$-bit integers can be carried out using three multiplications of n-bit integers, together with additions, subtractions, and shifts. This shows that if $f(n)$ is the total number of bit operations needed to multiply two n-bit integers, then

$$f(2n) = 3f(n) + Cn.$$

The reasoning behind this equation is as follows. The three multiplications of n-bit integers are carried out using $3f(n)$-bit operations. Each of the additions, subtractions, and shifts uses a constant multiple of n-bit operations, and Cn represents the total number of bit operations used by these operations. ◀

EXAMPLE 5 **Fast Matrix Multiplication** In Example 5 of Section 3.8 we showed that multiplying two $n \times n$ matrices using the definition of matrix multiplication required n^3 multiplications and $n^2(n-1)$ additions. Consequently, computing the product of two $n \times n$ matrices in this way requires $O(n^3)$ operations (multiplications and additions). Surprisingly, there are more efficient divide-and-conquer algorithms for multiplying two $n \times n$ matrices. Such an algorithm, invented by V. Strassen in 1969, reduces the multiplication of two $n \times n$ matrices, when n is even, to seven multiplications of two $(n/2) \times (n/2)$ matrices and 15 additions of $(n/2) \times (n/2)$ matrices. (See [CoLeRiSt01] for the details of this algorithm.) Hence, if $f(n)$ is the number of operations (multiplications and additions) used, it follows that

Links

$$f(n) = 7f(n/2) + 15n^2/4$$

when n is even. ◀

As Examples 1–5 show, recurrence relations of the form $f(n) = af(n/b) + g(n)$ arise in many different situations. It is possible to derive estimates of the size of functions that satisfy such recurrence relations. Suppose that f satisfies this recurrence relation whenever n is divisible by b. Let $n = b^k$, where k is a positive integer. Then

$$\begin{aligned} f(n) &= af(n/b) + g(n) \\ &= a^2 f(n/b^2) + ag(n/b) + g(n) \\ &= a^3 f(n/b^3) + a^2 g(n/b^2) + ag(n/b) + g(n) \\ &\vdots \\ &= a^k f(n/b^k) + \sum_{j=0}^{k-1} a^j g(n/b^j). \end{aligned}$$

Because $n/b^k = 1$, it follows that

$$f(n) = a^k f(1) + \sum_{j=0}^{k-1} a^j g(n/b^j).$$

We can use this equation for $f(n)$ to estimate the size of functions that satisfy divide-and-conquer relations.

THEOREM 1 Let f be an increasing function that satisfies the recurrence relation

$$f(n) = af(n/b) + c$$

whenever n is divisible by b, where $a \geq 1$, b is an integer greater than 1, and c is a positive real number. Then

$$f(n) \text{ is } \begin{cases} O(n^{\log_b a}) \text{ if } a > 1, \\ O(\log n) \text{ if } a = 1. \end{cases}$$

Furthermore, when $n = b^k$, where k is a positive integer,

$$f(n) = C_1 n^{\log_b a} + C_2,$$

where $C_1 = f(1) + c/(a-1)$ and $C_2 = -c/(a-1)$.

Proof: First let $n = b^k$. From the expression for $f(n)$ obtained in the discussion preceding the theorem, with $g(n) = c$, we have

$$f(n) = a^k f(1) + \sum_{j=0}^{k-1} a^j c = a^k f(1) + c \sum_{j=0}^{k-1} a^j.$$

First consider the case when $a = 1$. Then

$$f(n) = f(1) + ck.$$

Because $n = b^k$, we have $k = \log_b n$. Hence,

$$f(n) = f(1) + c \log_b n.$$

When n is not a power of b, we have $b^k < n < b^{k+1}$, for a positive integer k. Because f is increasing, it follows that $f(n) \leq f(b^{k+1}) = f(1) + c(k+1) = (f(1) + c) + ck \leq (f(1) + c) + c \log_b n$. Therefore, in both cases, $f(n)$ is $O(\log n)$ when $a = 1$.

Now suppose that $a > 1$. First assume that $n = b^k$, where k is a positive integer. From the formula for the sum of terms of a geometric progression (Theorem 1 in Section 2.4), it follows that

$$\begin{aligned} f(n) &= a^k f(1) + c(a^k - 1)/(a-1) \\ &= a^k[f(1) + c/(a-1)] - c/(a-1) \\ &= C_1 n^{\log_b a} + C_2, \end{aligned}$$

because $a^k = a^{\log_b n} = n^{\log_b a}$ (see Exercise 4 in Appendix 2), where $C_1 = f(1) + c/(a-1)$ and $C_2 = -c/(a-1)$.

Now suppose that n is not a power of b. Then $b^k < n < b^{k+1}$, where k is a nonnegative integer. Because f is increasing,

$$\begin{aligned} f(n) &\le f(b^{k+1}) = C_1 a^{k+1} + C_2 \\ &\le (C_1 a) a^{\log_b n} + C_2 \\ &\le (C_1 a) n^{\log_b a} + C_2, \end{aligned}$$

because $k \le \log_b n < k + 1$.

Hence, we have $f(n)$ is $O(n^{\log_b a})$. ◁

Examples 6–9 illustrate how Theorem 1 is used.

EXAMPLE 6 Let $f(n) = 5f(n/2) + 3$ and $f(1) = 7$. Find $f(2^k)$, where k is a positive integer. Also, estimate $f(n)$ if f is an increasing function.

Extra Examples

Solution: From the proof of Theorem 1, with $a = 5$, $b = 2$, and $c = 3$, we see that if $n = 2^k$, then

$$\begin{aligned} f(n) &= a^k[f(1) + c/(a-1)] + [-c/(a-1)] \\ &= 5^k[7 + (3/4)] - 3/4 \\ &= 5^k(31/4) - 3/4. \end{aligned}$$

Also, if $f(n)$ is increasing, Theorem 1 shows that $f(n)$ is $O(n^{\log_b a}) = O(n^{\log 5})$. ◀

We can use Theorem 1 to estimate the computational complexity of the binary search algorithm and the algorithm given in Example 2 for locating the minimum and maximum of a sequence.

EXAMPLE 7 Estimate the number of comparisons used by a binary search.

Solution: In Example 1 it was shown that $f(n) = f(n/2) + 2$ when n is even, where f is the number of comparisons required to perform a binary search on a sequence of size n. Hence, from Theorem 1, it follows that $f(n)$ is $O(\log n)$. ◀

EXAMPLE 8 Estimate the number of comparisons used to locate the maximum and minimum elements in a sequence using the algorithm given in Example 2.

Solution: In Example 2 we showed that $f(n) = 2f(n/2) + 2$, when n is even, where f is the number of comparisons needed by this algorithm. Hence, from Theorem 1, it follows that $f(n)$ is $O(n^{\log 2}) = O(n)$. ◀

We now state a more general, and more complicated, theorem, which has Theorem 1 as a special case. This theorem (or more powerful versions, including big-Theta estimates) is sometimes known as the Master Theorem because it is useful in analyzing the complexity of many important divide-and-conquer algorithms.

THEOREM 2 **MASTER THEOREM** Let f be an increasing function that satisfies the recurrence relation

$$f(n) = af(n/b) + cn^d$$

whenever $n = b^k$, where k is a positive integer, $a \geq 1$, b is an integer greater than 1, and c and d are real numbers with c positive and d nonnegative. Then

$$f(n) \text{ is } \begin{cases} O(n^d) & \text{if } a < b^d, \\ O(n^d \log n) & \text{if } a = b^d, \\ O(n^{\log_b a}) & \text{if } a > b^d. \end{cases}$$

The proof of Theorem 2 is left for the reader as Exercises 29–33 at the end of this section.

EXAMPLE 9 **Complexity of Merge Sort** In Example 3 we explained that the number of comparisons used by the merge sort to sort a list of n elements is less than $M(n)$, where $M(n) = 2M(n/2) + n$. By the Master Theorem (Theorem 2) we find that $M(n)$ is $O(n \log n)$, which agrees with the estimate found in Section 4.4. ◀

EXAMPLE 10 Estimate the number of bit operations needed to multiply two n-bit integers using the fast multiplication algorithm described in Example 4.

Solution: Example 4 shows that $f(n) = 3f(n/2) + Cn$, when n is even, where $f(n)$ is the number of bit operations required to multiply two n-bit integers using the fast multiplication algorithm. Hence, from the Master Theorem (Theorem 2), it follows that $f(n)$ is $O(n^{\log 3})$. Note that $\log 3 \sim 1.6$. Because the conventional algorithm for multiplication uses $O(n^2)$ bit operations, the fast multiplication algorithm is a substantial improvement over the conventional algorithm in terms of time complexity for sufficiently large integers, including large integers that occur in practical applications. ◀

EXAMPLE 11 Estimate the number of multiplications and additions required to multiply two $n \times n$ matrices using the matrix multiplication algorithm referred to in Example 5.

Solution: Let $f(n)$ denote the number of additions and multiplications used by the algorithm mentioned in Example 5 to multiply two $n \times n$ matrices. We have $f(n) = 7f(n/2) + 15n^2/4$, when n is even. Hence, from the Master Theorem (Theorem 2), it follows that $f(n)$ is $O(n^{\log 7})$. Note that $\log 7 \sim 2.8$. Because the conventional algorithm for multiplying two $n \times n$ matrices uses $O(n^3)$ additions and multiplications, it follows that for sufficiently large integers n, including those that occur in many practical applications, this algorithm is substantially more efficient in time complexity than the conventional algorithm. ◀

THE CLOSEST-PAIR PROBLEM We conclude this section by introducing a divide-and-conquer algorithm from computational geometry, the part of discrete mathematics devoted to algorithms that solve geometric problems.

EXAMPLE 12 **The Closest-Pair Problem** Consider the problem of determining the closest pair of points in a set of n points $(x_1, y_1), \ldots, (x_n, y_n)$ in the plane, where the distance between two points (x_i, y_i) and (x_j, y_j) is the usual Euclidean distance $\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. This problem arises in

Links

In this illustration the problem of finding the closest pair in a set of 16 points is reduced to two problems of finding the closest pair in a set of eight points *and* the problem of determining whether there are points closer than $d = \min(d_L, d_R)$ within the strip of width $2d$ centered at ℓ.

FIGURE 1 The Recursive Step of the Algorithm for Solving the Closest-Pair Problem.

many applications such as determining the closest pair of airplanes in the air space at a particular altitude being managed by an air traffic controller. How can this closest pair of points be found in an efficient way?

Solution: To solve this problem we can first determine the distance between every pair of points and then find the smallest of these distances. However, this approach requires $O(n^2)$ computations of distances and comparisons because there are $C(n, 2) = n(n-1)/2$ pairs of points. Surprisingly, there is an elegant divide-and-conquer algorithm that can solve the closest-pair problem for n points using $O(n \log n)$ computations of distances and comparisons. The algorithm we describe here is due to Michael Samos (see [PrSa85]).

For simplicity, we assume that $n = 2^k$, where k is a positive integer. (We avoid some technical considerations that are needed when n is not a power of 2.) When $n = 2$, we have only one pair of points; the distance between these two points is the minimum distance. At the start of the algorithm we use the merge sort twice, once to sort the points in order of increasing x coordinates, and once to sort the points in order of increasing y coordinates. Each of these sorts requires $O(n \log n)$ operations. We will use these sorted lists in each recursive step.

The recursive part of the algorithm divides the problem into two subproblems, each involving half as many points. Using the sorted list of the points by their x coordinates, we construct a vertical line ℓ dividing the n points into two parts, a left part and a right part of equal size, each containing $n/2$ points, as shown in Figure 1. (If any points fall on the dividing line ℓ, we divide them among the two parts if necessary.) At subsequent steps of the recursion we need not sort on x coordinates again, because we can select the corresponding sorted subset of all the points. This selection is a task that can be done with $O(n)$ comparisons.

There are three possibilities concerning the positions of the closest points: (1) they are both in the left region L, (2) they are both in the right region R, or (3) one point is in the left region and the other is in the right region. Apply the algorithm recursively to compute d_L and d_R, where d_L is the minimum distance between points in the left region and d_R is the minimum distance between points in the right region. Let $d = \min(d_L, d_R)$. To successfully divide the problem of finding the closest two points in the original set into the two problems of finding the shortest distances between points in the two regions separately, we have to handle the conquer part of the algorithm, which requires that we consider the case where the closest points lie in different regions, that is, one point is in L and the other in R. Because there is a pair of points at distance d where both points lie in R or both points lie in L, for the closest points to lie in different regions requires that they must be a distance less than d apart.

At most eight points, including p, can lie in or on the $2d \times d$ rectangle centered at ℓ because at most one point can lie in or on each of the eight $(d/2) \times (d/2)$ squares.

FIGURE 2 Showing That There Are at Most Seven Other Points to Consider for Each Point in the Strip.

For a point in the left region and a point in the right region to lie at a distance less than d apart, these points must lie in the vertical strip of width $2d$ that has the line ℓ as its center. (Otherwise, the distance between these points is greater than the difference in their x coordinates, which exceeds d.) To examine the points within this strip, we sort the points so that they are listed in order of increasing y coordinates, using the sorted list of the points by their y coordinates. At each recursive step, we form a subset of the points in the region sorted by their y coordinates from the already sorted set of all points sorted by their y coordinates, which can be done with $O(n)$ comparisons.

Beginning with a point in the strip with the smallest y coordinate, we successively examine each point in the strip, computing the distance between this point and all other points in the strip that have larger y coordinates that could lie at a distance less than d from this point. Note that to examine a point p, we need only consider the distances between p and points in the set that lie within the rectangle of height d and width $2d$ with p on its base and with vertical sides at distance d from ℓ.

We can show that there are at most eight points from the set, including p, in or on this $2d \times d$ rectangle. To see this, note that there can be at most one point in each of the eight $d/2 \times d/2$ squares shown in Figure 2. This follows because the farthest apart points can be on or within one of these squares is the diagonal length $d/\sqrt{2}$ (which can be found using the Pythagorean Theorem), which is less than d, and each of these $d/2 \times d/2$ squares lies entirely within the left region or the right region. This means that at this stage we need only compare at most seven distances, the distances between p and the seven or fewer other points in or on the rectangle, with d.

Because the total number of points in the strip of width $2d$ does not exceed n (the total number of points in the set), at most $7n$ distances need to be compared with d to find the minimum distance between points. That is, there are only $7n$ possible distances that could be less than d. Consequently, once the merge sort has been used to sort the pairs according to their x coordinates and according to their y coordinates, we find that the increasing function $f(n)$ satisfying the recurrence relation

$$f(n) = 2f(n/2) + 7n,$$

where $f(2) = 1$, exceeds the number of comparisons needed to solve the closest-pair problem for n points. By the Master Theorem (Theorem 2), it follows that $f(n)$ is $O(n \log n)$. The two sorts of points by their x coordinates and by their y coordinates each can be done using $O(n \log n)$ comparisons, by using the merge sort, and the sorted subsets of these coordinates at each of the $O(\log n)$ steps of the algorithm can be done using $O(n)$ comparisons each. Thus, we find that the closest-pair problem can be solved using $O(n \log n)$ comparisons. ◀

Exercises

1. How many comparisons are needed for a binary search in a set of 64 elements?
2. How many comparisons are needed to locate the maximum and minimum elements in a sequence with 128 elements using the algorithm in Example 2?
3. Multiply $(1110)_2$ and $(1010)_2$ using the fast multiplication algorithm.
4. Express the fast multiplication algorithm in pseudocode.
5. Determine a value for the constant C in Example 4 and use it to estimate the number of bit operations needed to multiply two 64-bit integers using the fast multiplication algorithm.
6. How many operations are needed to multiply two 32×32 matrices using the algorithm referred to in Example 4?
7. Suppose that $f(n) = f(n/3) + 1$ when n is divisible by 3, and $f(1) = 1$. Find
   a) $f(3)$. **b)** $f(27)$. **c)** $f(729)$.
8. Suppose that $f(n) = 2f(n/2) + 3$ when n is even, and $f(1) = 5$. Find
   a) $f(2)$. **b)** $f(8)$. **c)** $f(64)$. **d)** $f(1024)$.
9. Suppose that $f(n) = f(n/5) + 3n^2$ when n is divisible by 5, and $f(1) = 4$. Find
   a) $f(5)$. **b)** $f(125)$. **c)** $f(3125)$.
10. Find $f(n)$ when $n = 2^k$, where f satisfies the recurrence relation $f(n) = f(n/2) + 1$ with $f(1) = 1$.
11. Estimate the size of f in Exercise 10 if f is an increasing function.
12. Find $f(n)$ when $n = 3^k$, where f satisfies the recurrence relation $f(n) = 2f(n/3) + 4$ with $f(1) = 1$.
13. Estimate the size of f in Exercise 12 if f is an increasing function.
14. Suppose that there are $n = 2^k$ teams in an elimination tournament, where there are $n/2$ games in the first round, with the $n/2 = 2^{k-1}$ winners playing in the second round, and so on. Develop a recurrence relation for the number of rounds in the tournament.
15. How many rounds are in the elimination tournament described in Exercise 14 when there are 32 teams?
16. Solve the recurrence relation for the number of rounds in the tournament described in Exercise 14.
17. Suppose that the votes of n people for different candidates (where there can be more than two candidates) for a particular office are the elements of a sequence. A person wins the election if this person receives a majority of the votes.
    a) Devise a divide-and-conquer algorithm that determines whether a candidate received a majority and, if so, determine who this candidate is. [*Hint:* Assume that n is even and split the sequence of votes into two sequences, each with $n/2$ elements. Note that a candidate could not have received a majority of votes without receiving a majority of votes in at least one of the two halves.]
    b) Use the Master Theorem to estimate the number of comparisons needed by the algorithm you devised in part (a).
18. Suppose that each person in a group of n people votes for exactly two people from a slate of candidates to fill two positions on a committee. The top two finishers both win positions as long as each receives more than $n/2$ votes.
    a) Devise a divide-and-conquer algorithm that determines whether the two candidates who received the most votes each received at least $n/2$ votes and, if so, determine who these two candidates are.
    b) Use the Master Theorem to estimate the number of comparisons needed by the algorithm you devised in part (a).
19. **a)** Set up a divide-and-conquer recurrence relation for the number of multiplications required to compute x^n, where x is a real number and n is a positive integer, using the recursive algorithm from Exercise 26 in Section 4.4.
    b) Use the recurrence relation you found in part (a) to construct a big-O estimate for the number of multiplications used to compute x^n using the recursive algorithm.
20. **a)** Set up a divide-and-conquer recurrence relation for the number of modular multiplications required to compute a^n **mod** m, where a, m, and n are positive integers, using the recursive algorithms from Example 3 in Section 4.4.
    b) Use the recurrence relation you found in part (a) to construct a big-O estimate for the number of modular multiplications used to compute a^n **mod** m using the recursive algorithm.

21. Suppose that the function f satisfies the recurrence relation $f(n) = 2f(\sqrt{n}) + 1$ whenever n is a perfect square greater than 1 and $f(2) = 1$.

a) Find $f(16)$.

b) Find a big-O estimate for $f(n)$. [*Hint:* Make the substitution $m = \log n$.]

22. Suppose that the function f satisfies the recurrence relation $f(n) = 2f(\sqrt{n}) + \log n$ whenever n is a perfect square greater than 1 and $f(2) = 1$.

a) Find $f(16)$.

b) Find a big-O estimate for $f(n)$. [*Hint:* Make the substitution $m = \log n$.]

****23.** This exercise deals with the problem of finding the largest sum of consecutive terms of a sequence of n real numbers. When all terms are positive, the sum of all terms provides the answer, but the situation is more complicated when some terms are negative. For example, the maximum sum of consecutive terms of the sequence $-2, 3, -1, 6, -7, 4$ is $3 + (-1) + 6 = 8$. (This exercise is based on [Be86].)

a) Use pseudocode to describe an algorithm that solves this problem by finding the sums of consecutive terms starting with the first term, the sums of consecutive terms starting with the second term, and so on, keeping track of the maximum sum found so far as the algorithm proceeds.

b) Determine the computational complexity of the algorithm in part (a) in terms of the number of sums computed and the number of comparisons made.

c) Devise a divide-and-conquer algorithm to solve this problem. [*Hint:* Assume that there are an even number of terms in the sequence and split the sequence into two halves. Explain how to handle the case when the maximum sum of consecutive terms includes terms in both halves.]

d) Use the algorithm from part (c) to find the maximum sum of consecutive terms of each of the sequences: $-2, 4, -1, 3, 5, -6, 1, 2$; $4, 1, -3, 7, -1, -5, 3, -2$; and $-1, 6, 3, -4, -5, 8, -1, 7$.

e) Find a recurrence relation for the number of sums and comparisons used by the divide-and-conquer algorithm from part (c).

f) Use the Master Theorem to estimate the computational complexity of the divide-and-conquer algorithm. How does it compare in terms of computational complexity with the algorithm from part (a)?

24. Apply the algorithm described in Example 12 for finding the closest pair of points, using the Euclidean distance between points, to find the closest pair of the points $(1, 3)$, $(1, 7)$, $(2, 4)$, $(2, 9)$, $(3, 1)$, $(3, 5)$, $(4, 3)$, and $(4, 7)$.

25. Apply the algorithm described in Example 12 for finding the closest pair of points, using the Euclidean distance between points, to find the closest pair of the points $(1, 2)$, $(1, 6)$, $(2, 4)$, $(2, 8)$, $(3, 1)$, $(3, 6)$, $(3, 10)$, $(4, 3)$, $(5, 1)$, $(5, 5)$, $(5, 9)$, $(6, 7)$, $(7, 1)$, $(7, 4)$, $(7, 9)$, and $(8, 6)$.

***26.** Use pseudocode to describe the recursive algorithm for solving the closest-pair problem as described in Example 12.

27. Construct a variation of the algorithm described in Example 12 along with justifications of the steps used by the algorithm to find the smallest distance between two points if the distance between two points is defined to be $d((x_i, y_i), (x_j, y_j)) = \max(|x_i - x_j|, |y_i - y_j|)$.

***28.** Suppose someone picks a number x from a set of n numbers. A second person tries to guess the number by successively selecting subsets of the n numbers and asking the first person whether x is in each set. The first person answers either "yes" or "no." When the first person answers each query truthfully, we can find x using $\log n$ queries by successively splitting the sets used in each query in half. Ulam's problem, proposed by Stanislaw Ulam in 1976, asks for the number of queries required to find x, supposing that the first person is allowed to lie exactly once.

a) Show that by asking each question twice, given a number x and a set with n elements, and asking one more question when we find the lie, Ulam's problem can be solved using $2 \log n + 1$ queries.

b) Show that by dividing the initial set of n elements into four parts, each with $n/4$ elements, 1/4 of the elements can be eliminated using two queries. [*Hint:* Use two queries, where each of the queries asks whether the element is in the union of two of the subsets with $n/4$ elements and where one of the subsets of $n/4$ elements is used in both queries.]

c) Show from part (b) that if $f(n)$ equals the number of queries used to solve Ulam's problem using the method from part (b) and n is divisible by 4, then $f(n) = f(3n/4) + 2$.

d) Solve the recurrence relation in part (c) for $f(n)$.

e) Is the naive way to solve Ulam's problem by asking each question twice or the divide-and-conquer method based on part (b) more efficient? The most efficient way to solve Ulam's problem has been determined by A. Pelc [Pe87].

In Exercises 29–33, assume that f is an increasing function satisfying the recurrence relation $f(n) = af(n/b) + cn^d$, where $a \geq 1$, b is an integer greater than 1, and c and d are positive real numbers. These exercises supply a proof of Theorem 2.

***29.** Show that if $a = b^d$ and n is a power of b, then $f(n) = f(1)n^d + cn^d \log_b n$.

30. Use Exercise 29 to show that if $a = b^d$, then $f(n)$ is $O(n^d \log n)$.

***31.** Show that if $a \neq b^d$ and n is a power of b, then $f(n) = C_1 n^d + C_2 n^{\log_b a}$, where $C_1 = b^d c/(b^d - a)$ and $C_2 = f(1) + b^d c/(a - b^d)$.

32. Use Exercise 31 to show that if $a < b^d$, then $f(n)$ is $O(n^d)$.

33. Use Exercise 31 to show that if $a > b^d$, then $f(n)$ is $O(n^{\log_b a})$.

34. Find $f(n)$ when $n = 4^k$, where f satisfies the recurrence relation $f(n) = 5f(n/4) + 6n$, with $f(1) = 1$.
35. Estimate the size of f in Exercise 34 if f is an increasing function.
36. Find $f(n)$ when $n = 2^k$, where f satisfies the recurrence relation $f(n) = 8f(n/2) + n^2$ with $f(1) = 1$.
37. Estimate the size of f in Exercise 36 if f is an increasing function.

7.4 Generating Functions

Introduction

Links

Generating functions are used to represent sequences efficiently by coding the terms of a sequence as coefficients of powers of a variable x in a formal power series. Generating functions can be used to solve many types of counting problems, such as the number of ways to select or distribute objects of different kinds, subject to a variety of constraints, and the number of ways to make change for a dollar using coins of different denominations. Generating functions can be used to solve recurrence relations by translating a recurrence relation for the terms of a sequence into an equation involving a generating function. This equation can then be solved to find a closed form for the generating function. From this closed form, the coefficients of the power series for the generating function can be found, solving the original recurrence relation. Generating functions can also be used to prove combinatorial identities by taking advantage of relatively simple relationships between functions that can be translated into identities involving the terms of sequences. Generating functions are a helpful tool for studying many properties of sequences besides those described in this section, such as their use for establishing asymptotic formulae for the terms of a sequence.

We begin with the definition of the generating function for a sequence.

DEFINITION 1 The *generating function for the sequence* $a_0, a_1, \ldots, a_k, \ldots$ of real numbers is the infinite series

$$G(x) = a_0 + a_1x + \cdots + a_kx^k + \cdots = \sum_{k=0}^{\infty} a_kx^k.$$

Remark: The generating function for $\{a_k\}$ given in Definition 1 is sometimes called the **ordinary generating function** of $\{a_k\}$ to distinguish it from other types of generating functions for this sequence.

EXAMPLE 1 The generating functions for the sequences $\{a_k\}$ with $a_k = 3$, $a_k = k + 1$, and $a_k = 2^k$ are $\sum_{k=0}^{\infty} 3x^k$, $\sum_{k=0}^{\infty}(k + 1)x^k$, and $\sum_{k=0}^{\infty} 2^kx^k$, respectively. ◀

Extra Examples

We can define generating functions for finite sequences of real numbers by extending a finite sequence $a_0, a_1, \ldots, a_n$ into an infinite sequence by setting $a_{n+1} = 0$, $a_{n+2} = 0$, and so on. The generating function $G(x)$ of this infinite sequence $\{a_n\}$ is a polynomial of degree n because no terms of the form a_jx^j with $j > n$ occur, that is,

$$G(x) = a_0 + a_1x + \cdots + a_nx^n.$$

EXAMPLE 2 What is the generating function for the sequence 1, 1, 1, 1, 1, 1?

Solution: The generating function of 1, 1, 1, 1, 1, 1 is

$$1 + x + x^2 + x^3 + x^4 + x^5.$$

By Theorem 1 of Section 2.4 we have

$$(x^6 - 1)/(x - 1) = 1 + x + x^2 + x^3 + x^4 + x^5$$

when $x \neq 1$. Consequently, $G(x) = (x^6 - 1)/(x - 1)$ is the generating function of the sequence 1, 1, 1, 1, 1, 1. [Because the powers of x are only place holders for the terms of the sequence in a generating function, we do not need to worry that $G(1)$ is undefined.] ◀

EXAMPLE 3 Let m be a positive integer. Let $a_k = C(m, k)$, for $k = 0, 1, 2, \ldots, m$. What is the generating function for the sequence $a_0, a_1, \ldots, a_m$?

Solution: The generating function for this sequence is

$$G(x) = C(m, 0) + C(m, 1)x + C(m, 2)x^2 + \cdots + C(m, m)x^m.$$

The Binomial Theorem shows that $G(x) = (1 + x)^m$. ◀

Useful Facts About Power Series

When generating functions are used to solve counting problems, they are usually considered to be **formal power series.** Questions about the convergence of these series are ignored. However, to apply some results from calculus, it is sometimes important to consider for which x the power series converges. The fact that a function has a unique power series around $x = 0$ will also be important. Generally, however, we will not be concerned with questions of convergence or the uniqueness of power series in our discussions. Readers familiar with calculus can consult textbooks on this subject for details about power series, including the convergence of the series we consider here.

We will now state some important facts about infinite series used when working with generating functions. A discussion of these and related results can be found in calculus texts.

EXAMPLE 4 The function $f(x) = 1/(1 - x)$ is the generating function of the sequence 1, 1, 1, 1, ..., because

$$1/(1 - x) = 1 + x + x^2 + \cdots$$

for $|x| < 1$. ◀

EXAMPLE 5 The function $f(x) = 1/(1 - ax)$ is the generating function of the sequence $1, a, a^2, a^3, \ldots$, because

$$1/(1 - ax) = 1 + ax + a^2x^2 + \cdots$$

when $|ax| < 1$, or equivalently, for $|x| < 1/|a|$ for $a \neq 0$. ◀

We also will need some results on how to add and how to multiply two generating functions. Proofs of these results can be found in calculus texts.

THEOREM 1 Let $f(x) = \sum_{k=0}^{\infty} a_k x^k$ and $g(x) = \sum_{k=0}^{\infty} b_k x^k$. Then

$$f(x) + g(x) = \sum_{k=0}^{\infty} (a_k + b_k) x^k \quad \text{and} \quad f(x)g(x) = \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} a_j b_{k-j} \right) x^k.$$

Remark: Theorem 1 is valid only for power series that converge in an interval, as all series considered in this section do. However, the theory of generating functions is not limited to such series. In the case of series that do not converge, the statements in Theorem 1 can be taken as definitions of addition and multiplication of generating functions.

We will illustrate how Theorem 1 can be used with Example 6.

EXAMPLE 6 Let $f(x) = 1/(1-x)^2$. Use Example 4 to find the coefficients $a_0, a_1, a_2, \ldots$ in the expansion $f(x) = \sum_{k=0}^{\infty} a_k x^k$.

Solution: From Example 4 we see that

$$1/(1-x) = 1 + x + x^2 + x^3 + \cdots.$$

Hence, from Theorem 1, we have

$$1/(1-x)^2 = \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} 1 \right) x^k = \sum_{k=0}^{\infty} (k+1) x^k.$$

◀

Remark: This result also can be derived from Example 4 by differentiation. Taking derivatives is a useful technique for producing new identities from existing identities for generating functions.

To use generating functions to solve many important counting problems, we will need to apply the Binomial Theorem for exponents that are not positive integers. Before we state an extended version of the Binomial Theorem, we need to define extended binomial coefficients.

DEFINITION 2 Let u be a real number and k a nonnegative integer. Then the *extended binomial coefficient* $\binom{u}{k}$ is defined by

$$\binom{u}{k} = \begin{cases} u(u-1)\cdots(u-k+1)/k! & \text{if } k > 0, \\ 1 & \text{if } k = 0. \end{cases}$$

EXAMPLE 7 Find the values of the extended binomial coefficients $\binom{-2}{3}$ and $\binom{1/2}{3}$.

Solution: Taking $u = -2$ and $k = 3$ in Definition 2 gives us

$$\binom{-2}{3} = \frac{(-2)(-3)(-4)}{3!} = -4.$$

Similarly, taking $u = 1/2$ and $k = 3$ gives us

$$\begin{aligned}\binom{1/2}{3} &= \frac{(1/2)(1/2-1)(1/2-2)}{3!}\\ &= (1/2)(-1/2)(-3/2)/6\\ &= 1/16.\end{aligned}$$

◀

Example 8 provides a useful formula for extended binomial coefficients when the top parameter is a negative integer. It will be useful in our subsequent discussions.

EXAMPLE 8 When the top parameter is a negative integer, the extended binomial coefficient can be expressed in terms of an ordinary binomial coefficient. To see that this is the case, note that

$$\begin{aligned}\binom{-n}{r} &= \frac{(-n)(-n-1)\cdots(-n-r+1)}{r!} && \text{by definition of extended binomial coefficient}\\ &= \frac{(-1)^r n(n+1)\cdots(n+r-1)}{r!} && \text{factoring out } -1 \text{ from each term in the numerator}\\ &= \frac{(-1)^r (n+r-1)(n+r-2)\cdots n}{r!} && \text{by the commutative law for multiplication}\\ &= \frac{(-1)^r (n+r-1)!}{r!(n-1)!} && \text{multiplying both the numerator and denominator by } (n-1)!\\ &= (-1)^r \binom{n+r-1}{r} && \text{by the definition of binomial coefficients}\\ &= (-1)^r C(n+r-1, r). && \text{using alternative notation for binomial coefficients}\end{aligned}$$

◀

We now state the extended Binomial Theorem.

THEOREM 2 **THE EXTENDED BINOMIAL THEOREM** Let x be a real number with $|x| < 1$ and let u be a real number. Then

$$(1+x)^u = \sum_{k=0}^{\infty} \binom{u}{k} x^k.$$

Theorem 2 can be proved using the theory of the Maclaurin series. We leave its proof to the reader with a familiarity with this part of calculus.

Remark: When u is a positive integer, the extended Binomial Theorem reduces to the Binomial Theorem presented in Section 5.4, because in that case $\binom{u}{k} = 0$ if $k > u$.

Example 9 illustrates the use of Theorem 2 when the exponent is a negative integer.

EXAMPLE 9 Find the generating functions for $(1+x)^{-n}$ and $(1-x)^{-n}$, where n is a positive integer, using the extended Binomial Theorem.

Solution: By the extended Binomial Theorem, it follows that

$$(1+x)^{-n} = \sum_{k=0}^{\infty} \binom{-n}{k} x^k.$$

Using Example 8, which provides a simple formula for $\binom{-n}{k}$, we obtain

$$(1+x)^{-n} = \sum_{k=0}^{\infty} (-1)^k C(n+k-1, k) x^k.$$

Replacing x by $-x$, we find that

$$(1-x)^{-n} = \sum_{k=0}^{\infty} C(n+k-1, k) x^k.$$

◀

Table 1 presents a useful summary of some generating functions that arise frequently.

Remark: Note that the second and third formulae in this table can be deduced from the first formula by substituting ax and x^r for x, respectively. Similarly, the sixth and seventh formulae can be deduced from the fifth formula using the same substitutions. The tenth and eleventh can be deduced from the ninth formula by substituting $-x$ and ax for x, respectively. Also, some of the formulae in this table can be derived from other formulae using methods from calculus (such as differentiation and integration). Students are encouraged to know the core formulae in this table (that is, formulae from which the others can be derived, perhaps the first, fourth, fifth, eighth, ninth, twelfth, and thirteenth formulae) and understand how to derive the other formulae from these core formulae.

Counting Problems and Generating Functions

Generating functions can be used to solve a wide variety of counting problems. In particular, they can be used to count the number of combinations of various types. In Chapter 5 we developed techniques to count the r-combinations from a set with n elements when repetition is allowed and additional constraints may exist. Such problems are equivalent to counting the solutions to equations of the form

$$e_1 + e_2 + \cdots + e_n = C,$$

where C is a constant and each e_i is a nonnegative integer that may be subject to a specified constraint. Generating functions can also be used to solve counting problems of this type, as Examples 10–12 show.

EXAMPLE 10 Find the number of solutions of

$$e_1 + e_2 + e_3 = 17,$$

where e_1, e_2, and e_3 are nonnegative integers with $2 \le e_1 \le 5$, $3 \le e_2 \le 6$, and $4 \le e_3 \le 7$.

TABLE 1 Useful Generating Functions.

$G(x)$	a_k
$(1+x)^n = \sum_{k=0}^{n} C(n,k)x^k = 1 + C(n,1)x + C(n,2)x^2 + \cdots + x^n$	$C(n,k)$
$(1+ax)^n = \sum_{k=0}^{n} C(n,k)a^k x^k = 1 + C(n,1)ax + C(n,2)a^2x^2 + \cdots + a^n x^n$	$C(n,k)a^k$
$(1+x^r)^n = \sum_{k=0}^{n} C(n,k)x^{rk} = 1 + C(n,1)x^r + C(n,2)x^{2r} + \cdots + x^{rn}$	$C(n,k/r)$ if $r \mid k$; 0 otherwise
$\frac{1-x^{n+1}}{1-x} = \sum_{k=0}^{n} x^k = 1 + x + x^2 + \cdots + x^n$	1 if $k \le n$; 0 otherwise
$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + \cdots$	1
$\frac{1}{1-ax} = \sum_{k=0}^{\infty} a^k x^k = 1 + ax + a^2x^2 + \cdots$	a^k
$\frac{1}{1-x^r} = \sum_{k=0}^{\infty} x^{rk} = 1 + x^r + x^{2r} + \cdots$	1 if $r \mid k$; 0 otherwise
$\frac{1}{(1-x)^2} = \sum_{k=0}^{\infty} (k+1)x^k = 1 + 2x + 3x^2 + \cdots$	$k+1$
$\frac{1}{(1-x)^n} = \sum_{k=0}^{\infty} C(n+k-1,k)x^k = 1 + C(n,1)x + C(n+1,2)x^2 + \cdots$	$C(n+k-1,k) = C(n+k-1,n-1)$
$\frac{1}{(1+x)^n} = \sum_{k=0}^{\infty} C(n+k-1,k)(-1)^k x^k = 1 - C(n,1)x + C(n+1,2)x^2 - \cdots$	$(-1)^k C(n+k-1,k) = (-1)^k C(n+k-1,n-1)$
$\frac{1}{(1-ax)^n} = \sum_{k=0}^{\infty} C(n+k-1,k)a^k x^k = 1 + C(n,1)ax + C(n+1,2)a^2x^2 + \cdots$	$C(n+k-1,k)a^k = C(n+k-1,n-1)a^k$
$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$	$1/k!$
$\ln(1+x) = \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{k} x^k = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$	$(-1)^{k+1}/k$

Note. The series for the last two generating functions can be found in most calculus books when power series are discussed.

Solution: The number of solutions with the indicated constraints is the coefficient of x^{17} in the expansion of

$$(x^2 + x^3 + x^4 + x^5)(x^3 + x^4 + x^5 + x^6)(x^4 + x^5 + x^6 + x^7).$$

This follows because we obtain a term equal to x^{17} in the product by picking a term in the first sum x^{e_1}, a term in the second sum x^{e_2}, and a term in the third sum x^{e_3}, where the exponents e_1, e_2, and e_3 satisfy the equation $e_1 + e_2 + e_3 = 17$ and the given constraints.

It is not hard to see that the coefficient of x^{17} in this product is 3. Hence, there are three solutions. (Note that the calculating of this coefficient involves about as much work as enumerating all the solutions of the equation with the given constraints. However, the method that this illustrates often can be used to solve wide classes of counting problems with special formulae, as we will see. Furthermore, a computer algebra system can be used to do such computations.) ◀

EXAMPLE 11 In how many different ways can eight identical cookies be distributed among three distinct children if each child receives at least two cookies and no more than four cookies?

Solution: Because each child receives at least two but no more than four cookies, for each child there is a factor equal to

$$(x^2 + x^3 + x^4)$$

in the generating function for the sequence $\{c_n\}$, where c_n is the number of ways to distribute n cookies. Because there are three children, this generating function is

$$(x^2 + x^3 + x^4)^3.$$

We need the coefficient of x^8 in this product. The reason is that the x^8 terms in the expansion correspond to the ways that three terms can be selected, with one from each factor, that have exponents adding up to 8. Furthermore, the exponents of the term from the first, second, and third factors are the numbers of cookies the first, second, and third children receive, respectively. Computation shows that this coefficient equals 6. Hence, there are six ways to distribute the cookies so that each child receives at least two, but no more than four, cookies. ◀

EXAMPLE 12 Use generating functions to determine the number of ways to insert tokens worth \$1, \$2, and \$5 into a vending machine to pay for an item that costs r dollars in both the cases when the order in which the tokens are inserted does not matter and when the order does matter. (For example, there are two ways to pay for an item that costs \$3 when the order in which the tokens are inserted does not matter: inserting three \$1 tokens or one \$1 token and a \$2 token. When the order matters, there are three ways: inserting three \$1 tokens, inserting a \$1 token and then a \$2 token, or inserting a \$2 token and then a \$1 token.)

Solution: Consider the case when the order in which the tokens are inserted does not matter. Here, all we care about is the number of each token used to produce a total of r dollars. Because we can use any number of \$1 tokens, any number of \$2 tokens, and any number of \$5 tokens, the answer is the coefficient of x^r in the generating function

$$(1 + x + x^2 + x^3 + \cdots)(1 + x^2 + x^4 + x^6 + \cdots)(1 + x^5 + x^{10} + x^{15} + \cdots).$$

(The first factor in this product represents the \$1 tokens used, the second the \$2 tokens used, and the third the \$5 tokens used.) For example, the number of ways to pay for an item costing \$7 using \$1, \$2, and \$5 tokens is given by the coefficient of x^7 in this expansion, which equals 6.

When the order in which the tokens are inserted matters, the number of ways to insert exactly n tokens to produce a total of r dollars is the coefficient of x^r in

$$(x + x^2 + x^5)^n,$$

because each of the r tokens may be a \$1 token, a \$2 token, or a \$5 token. Because any number of tokens may be inserted, the number of ways to produce r dollars using \$1, \$2, or \$5 tokens, when the order in which the tokens are inserted matters, is the coefficient of x^r in

$$\begin{aligned} 1 + (x + x^2 + x^5) + (x + x^2 + x^5)^2 + \cdots &= \frac{1}{1 - (x + x^2 + x^5)} \\ &= \frac{1}{1 - x - x^2 - x^5}, \end{aligned}$$

where we have added the number of ways to insert 0 tokens, 1 token, 2 tokens, 3 tokens, and so on, and where we have used the identity $1/(1 - x) = 1 + x + x^2 + \cdots$ with x replaced with $x + x^2 + x^5$. For example, the number of ways to pay for an item costing \$7 using \$1, \$2, and \$5 tokens, when the order in which the tokens are used matters, is the coefficient of x^7 in this expansion, which equals 26. [*Hint:* To see that this coefficient equals 26 requires the addition of the coefficients of x^7 in the expansions $(x + x^2 + x^5)^k$ for $2 \le k \le 7$. This can be done by hand with considerable computation, or a computer algebra system can be used.] ◀

Example 13 shows the versatility of generating functions when used to solve problems with differing assumptions.

EXAMPLE 13 Use generating functions to find the number of k-combinations of a set with n elements. Assume that the Binomial Theorem has already been established.

Solution: Each of the n elements in the set contributes the term $(1 + x)$ to the generating function $f(x) = \sum_{k=0}^{n} a_k x^k$. Here $f(x)$ is the generating function for $\{a_k\}$, where a_k represents the number of k-combinations of a set with n elements. Hence,

$$f(x) = (1 + x)^n.$$

But by the Binomial Theorem, we have

$$f(x) = \sum_{k=0}^{n} \binom{n}{k} x^k,$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Hence, $C(n, k)$, the number of k-combinations of a set with n elements, is

$$\frac{n!}{k!(n-k)!}.$$

◀

Remark: We proved the Binomial Theorem in Section 5.4 using the formula for the number of r-combinations of a set with n elements. This example shows that the Binomial Theorem, which can be proved by mathematical induction, can be used to derive the formula for the number of r-combinations of a set with n elements.

EXAMPLE 14 Use generating functions to find the number of r-combinations from a set with n elements when repetition of elements is allowed.

Solution: Let $G(x)$ be the generating function for the sequence $\{a_r\}$, where a_r equals the number of r-combinations of a set with n elements with repetitions allowed. That is, $G(x) = \sum_{r=0}^{\infty} a_r x^r$. Because we can select any number of a particular member of the set with n elements when we form an r-combination with repetition allowed, each of the n elements contributes $(1 + x + x^2 + x^3 + \cdots)$ to a product expansion for $G(x)$. Each element contributes this factor because it may be selected zero times, one time, two times, three times, and so on, when an r-combination is formed (with a total of r elements selected). Because there are n elements in the set and each contributes this same factor to $G(x)$, we have

$$G(x) = (1 + x + x^2 + \cdots)^n.$$

As long as $|x| < 1$, we have $1 + x + x^2 + \cdots = 1/(1 - x)$, so

$$G(x) = 1/(1 - x)^n = (1 - x)^{-n}.$$

Applying the extended Binomial Theorem (Theorem 2), it follows that

$$(1 - x)^{-n} = (1 + (-x))^{-n} = \sum_{r=0}^{\infty} \binom{-n}{r} (-x)^r.$$

The number of r-combinations of a set with n elements with repetitions allowed, when r is a positive integer, is the coefficient a_r of x^r in this sum. Consequently, using Example 8 we find that a_r equals

$$\begin{aligned}\binom{-n}{r}(-1)^r &= (-1)^r C(n + r - 1, r) \cdot (-1)^r \\ &= C(n + r - 1, r).\end{aligned}$$

◀

Note that the result in Example 14 is the same result we stated as Theorem 2 in Section 5.5.

EXAMPLE 15 Use generating functions to find the number of ways to select r objects of n different kinds if we must select at least one object of each kind.

Solution: Because we need to select at least one object of each kind, each of the n kinds of objects contributes the factor $(x + x^2 + x^3 + \cdots)$ to the generating function $G(x)$ for the sequence $\{a_r\}$, where a_r is the number of ways to select r objects of n different kinds if we need at least one object of each kind. Hence,

$$G(x) = (x + x^2 + x^3 + \cdots)^n = x^n(1 + x + x^2 + \cdots)^n = x^n/(1 - x)^n.$$

Using the extended Binomial Theorem and Example 8, we have

$$\begin{aligned}
G(x) &= x^n/(1-x)^n \\
&= x^n \cdot (1-x)^{-n} \\
&= x^n \sum_{r=0}^{\infty} \binom{-n}{r} (-x)^r \\
&= x^n \sum_{r=0}^{\infty} (-1)^r C(n+r-1, r)(-1)^r x^r \\
&= \sum_{r=0}^{\infty} C(n+r-1, r) x^{n+r} \\
&= \sum_{t=n}^{\infty} C(t-1, t-n) x^t \\
&= \sum_{r=n}^{\infty} C(r-1, r-n) x^r.
\end{aligned}$$

We have shifted the summation in the next-to-last equality by setting $t = n + r$ so that $t = n$ when $r = 0$ and $n + r - 1 = t - 1$, and then we replaced t by r as the index of summation in the last equality to return to our original notation. Hence, there are $C(r-1, r-n)$ ways to select r objects of n different kinds if we must select at least one object of each kind. ◀

Using Generating Functions to Solve Recurrence Relations

We can find the solution to a recurrence relation and its initial conditions by finding an explicit formula for the associated generating function. This is illustrated in Examples 16 and 17.

EXAMPLE 16 Solve the recurrence relation $a_k = 3a_{k-1}$ for $k = 1, 2, 3, \ldots$ and initial condition $a_0 = 2$.

Extra Examples

Solution: Let $G(x)$ be the generating function for the sequence $\{a_k\}$, that is, $G(x) = \sum_{k=0}^{\infty} a_k x^k$. First note that

$$xG(x) = \sum_{k=0}^{\infty} a_k x^{k+1} = \sum_{k=1}^{\infty} a_{k-1} x^k.$$

Using the recurrence relation, we see that

$$\begin{aligned}
G(x) - 3xG(x) &= \sum_{k=0}^{\infty} a_k x^k - 3 \sum_{k=1}^{\infty} a_{k-1} x^k \\
&= a_0 + \sum_{k=1}^{\infty} (a_k - 3a_{k-1}) x^k \\
&= 2,
\end{aligned}$$

because $a_0 = 2$ and $a_k = 3a_{k-1}$. Thus,

$$G(x) - 3xG(x) = (1 - 3x)G(x) = 2.$$

Solving for $G(x)$ shows that $G(x) = 2/(1 - 3x)$. Using the identity $1/(1 - ax) = \sum_{k=0}^{\infty} a^k x^k$, from Table 1, we have

$$G(x) = 2\sum_{k=0}^{\infty} 3^k x^k = \sum_{k=0}^{\infty} 2 \cdot 3^k x^k.$$

Consequently, $a_k = 2 \cdot 3^k$. ◀

EXAMPLE 17 Suppose that a valid codeword is an n-digit number in decimal notation containing an even number of 0s. Let a_n denote the number of valid codewords of length n. In Example 7 of Section 7.1 we showed that the sequence $\{a_n\}$ satisfies the recurrence relation

$$a_n = 8a_{n-1} + 10^{n-1}$$

and the initial condition $a_1 = 9$. Use generating functions to find an explicit formula for a_n.

Solution: To make our work with generating functions simpler, we extend this sequence by setting $a_0 = 1$; when we assign this value to a_0 and use the recurrence relation, we have $a_1 = 8a_0 + 10^0 = 8 + 1 = 9$, which is consistent with our original initial condition. (It also makes sense because there is one code word of length 0—the empty string.)

We multiply both sides of the recurrence relation by x^n to obtain

$$a_n x^n = 8a_{n-1}x^n + 10^{n-1}x^n.$$

Let $G(x) = \sum_{n=0}^{\infty} a_n x^n$ be the generating function of the sequence $a_0, a_1, a_2, \ldots$. We sum both sides of the last equation starting with $n = 1$, to find that

$$\begin{aligned}
G(x) - 1 &= \sum_{n=1}^{\infty} a_n x^n = \sum_{n=1}^{\infty}(8a_{n-1}x^n + 10^{n-1}x^n) \\
&= 8\sum_{n=1}^{\infty} a_{n-1}x^n + \sum_{n=1}^{\infty} 10^{n-1}x^n \\
&= 8x\sum_{n=1}^{\infty} a_{n-1}x^{n-1} + x\sum_{n=1}^{\infty} 10^{n-1}x^{n-1} \\
&= 8x\sum_{n=0}^{\infty} a_n x^n + x\sum_{n=0}^{\infty} 10^n x^n \\
&= 8xG(x) + x/(1 - 10x),
\end{aligned}$$

where we have used Example 5 to evaluate the second summation. Therefore, we have

$$G(x) - 1 = 8xG(x) + x/(1 - 10x).$$

Solving for $G(x)$ shows that

$$G(x) = \frac{1 - 9x}{(1 - 8x)(1 - 10x)}.$$

Expanding the right-hand side of this equation into partial fractions (as is done in the integration of rational functions studied in calculus) gives

$$G(x) = \frac{1}{2}\left(\frac{1}{1-8x} + \frac{1}{1-10x}\right).$$

Using Example 5 twice (once with $a = 8$ and once with $a = 10$) gives

$$\begin{aligned} G(x) &= \frac{1}{2}\left(\sum_{n=0}^{\infty} 8^n x^n + \sum_{n=0}^{\infty} 10^n x^n\right) \\ &= \sum_{n=0}^{\infty} \frac{1}{2}(8^n + 10^n)x^n. \end{aligned}$$

Consequently, we have shown that

$$a_n = \frac{1}{2}(8^n + 10^n).$$

◀

Proving Identities via Generating Functions

In Chapter 5 we saw how combinatorial identities could be established using combinatorial proofs. Here we will show that such identities, as well as identities for extended binomial coefficients, can be proved using generating functions. Sometimes the generating function approach is simpler than other approaches, especially when it is simpler to work with the closed form of a generating function than with the terms of the sequence themselves. We illustrate how generating functions can be used to prove identities with Example 18.

EXAMPLE 18 Use generating functions to show that

$$\sum_{k=0}^{n} C(n,k)^2 = C(2n,n)$$

whenever n is a positive integer.

Solution: First note that by the Binomial Theorem $C(2n, n)$ is the coefficient of x^n in $(1+x)^{2n}$. However, we also have

$$\begin{aligned} (1+x)^{2n} &= [(1+x)^n]^2 \\ &= [C(n,0) + C(n,1)x + C(n,2)x^2 + \cdots + C(n,n)x^n]^2. \end{aligned}$$

The coefficient of x^n in this expression is

$$C(n,0)C(n,n) + C(n,1)C(n,n-1) + C(n,2)C(n,n-2) + \cdots + C(n,n)C(n,0).$$

This equals $\sum_{k=0}^{n} C(n,k)^2$, because $C(n, n-k) = C(n,k)$. Because both $C(2n,n)$ and $\sum_{k=0}^{n} C(n,k)^2$ represent the coefficient of x^n in $(1+x)^{2n}$, they must be equal. ◀

Exercises 42 and 43 at the end of this section ask that Pascal's identity and Vandermonde's identity be proved using generating functions.

Exercises

1. Find the generating function for the finite sequence 2, 2, 2, 2, 2, 2.
2. Find the generating function for the finite sequence 1, 4, 16, 64, 256.

In Exercises 3–8, by a **closed form** we mean an algebraic expression not involving a summation over a range of values or the use of ellipses.

3. Find a closed form for the generating function for each of these sequences. (For each sequence, use the most obvious choice of a sequence that follows the pattern of the initial terms listed.)
   a) $0, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, \ldots$
   b) $0, 0, 0, 1, 1, 1, 1, 1, 1, \ldots$
   c) $0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, \ldots$
   d) $2, 4, 8, 16, 32, 64, 128, 256, \ldots$
   e) $\binom{7}{0}, \binom{7}{1}, \binom{7}{2}, \ldots, \binom{7}{7}, 0, 0, 0, 0, 0, \ldots$
   f) $2, -2, 2, -2, 2, -2, 2, -2, \ldots$
   g) $1, 1, 0, 1, 1, 1, 1, 1, 1, 1, \ldots$
   h) $0, 0, 0, 1, 2, 3, 4, \ldots$
4. Find a closed form for the generating function for each of these sequences. (Assume a general form for the terms of the sequence, using the most obvious choice of such a sequence.)
   a) $-1, -1, -1, -1, -1, -1, -1, 0, 0, 0, 0, 0, 0, \ldots$
   b) $1, 3, 9, 27, 81, 243, 729, \ldots$
   c) $0, 0, 3, -3, 3, -3, 3, -3, \ldots$
   d) $1, 2, 1, 1, 1, 1, 1, 1, 1, \ldots$
   e) $\binom{7}{0}, 2\binom{7}{1}, 2^2\binom{7}{2}, \ldots, 2^7\binom{7}{7}, 0, 0, 0, 0, \ldots$
   f) $-3, 3, -3, 3, -3, 3, \ldots$
   g) $0, 1, -2, 4, -8, 16, -32, 64, \ldots$
   h) $1, 0, 1, 0, 1, 0, 1, 0, \ldots$
5. Find a closed form for the generating function for the sequence $\{a_n\}$, where
   a) $a_n = 5$ for all $n = 0, 1, 2, \ldots$.
   b) $a_n = 3^n$ for all $n = 0, 1, 2, \ldots$.
   c) $a_n = 2$ for $n = 3, 4, 5, \ldots$ and $a_0 = a_1 = a_2 = 0$.
   d) $a_n = 2n + 3$ for all $n = 0, 1, 2, \ldots$.
   e) $a_n = \binom{8}{n}$ for all $n = 0, 1, 2, \ldots$.
   f) $a_n = \binom{n+4}{n}$ for all $n = 0, 1, 2, \ldots$.
6. Find a closed form for the generating function for the sequence $\{a_n\}$, where
   a) $a_n = -1$ for all $n = 0, 1, 2, \ldots$.
   b) $a_n = 2^n$ for $n = 1, 2, 3, 4, \ldots$ and $a_0 = 0$.
   c) $a_n = n - 1$ for $n = 0, 1, 2, \ldots$.
   d) $a_n = 1/(n+1)!$ for $n = 0, 1, 2, \ldots$.
   e) $a_n = \binom{n}{2}$ for $n = 0, 1, 2, \ldots$.
   f) $a_n = \binom{10}{n+1}$ for $n = 0, 1, 2, \ldots$.
7. For each of these generating functions, provide a closed formula for the sequence it determines.
   a) $(3x - 4)^3$ **b)** $(x^3 + 1)^3$
   c) $1/(1 - 5x)$ **d)** $x^3/(1 + 3x)$
   e) $x^2 + 3x + 7 + (1/(1 - x^2))$
   f) $(x^4/(1 - x^4)) - x^3 - x^2 - x - 1$
   g) $x^2/(1 - x)^2$ **h)** $2e^{2x}$
8. For each of these generating functions, provide a closed formula for the sequence it determines.
   a) $(x^2 + 1)^3$ **b)** $(3x - 1)^3$
   c) $1/(1 - 2x^2)$ **d)** $x^2/(1 - x)^3$
   e) $x - 1 + (1/(1 - 3x))$ **f)** $(1 + x^3)/(1 + x)^3$
   ***g)** $x/(1 + x + x^2)$ **h)** $e^{3x^2} - 1$
9. Find the coefficient of x^{10} in the power series of each of these functions.
   a) $(1 + x^5 + x^{10} + x^{15} + \cdots)^3$
   b) $(x^3 + x^4 + x^5 + x^6 + x^7 + \cdots)^3$
   c) $(x^4 + x^5 + x^6)(x^3 + x^4 + x^5 + x^6 + x^7)(1 + x + x^2 + x^3 + x^4 + \cdots)$
   d) $(x^2 + x^4 + x^6 + x^8 + \cdots)(x^3 + x^6 + x^9 + \cdots)(x^4 + x^8 + x^{12} + \cdots)$
   e) $(1 + x^2 + x^4 + x^6 + x^8 + \cdots)(1 + x^4 + x^8 + x^{12} + \cdots)(1 + x^6 + x^{12} + x^{18} + \cdots)$
10. Find the coefficient of x^9 in the power series of each of these functions.
    a) $(1 + x^3 + x^6 + x^9 + \cdots)^3$
    b) $(x^2 + x^3 + x^4 + x^5 + x^6 + \cdots)^3$
    c) $(x^3 + x^5 + x^6)(x^3 + x^4)(x + x^2 + x^3 + x^4 + \cdots)$
    d) $(x + x^4 + x^7 + x^{10} + \cdots)(x^2 + x^4 + x^6 + x^8 + \cdots)$
    e) $(1 + x + x^2)^3$
11. Find the coefficient of x^{10} in the power series of each of these functions.
    a) $1/(1 - 2x)$ **b)** $1/(1 + x)^2$
    c) $1/(1 - x)^3$ **d)** $1/(1 + 2x)^4$
    e) $x^4/(1 - 3x)^3$
12. Find the coefficient of x^{12} in the power series of each of these functions.
    a) $1/(1 + 3x)$ **b)** $1/(1 - 2x)^2$
    c) $1/(1 + x)^8$ **d)** $1/(1 - 4x)^3$
    e) $x^3/(1 + 4x)^2$
13. Use generating functions to determine the number of different ways 10 identical balloons can be given to four children if each child receives at least two balloons.
14. Use generating functions to determine the number of different ways 12 identical action figures can be given to five children so that each child receives at most three action figures.
15. Use generating functions to determine the number of different ways 15 identical stuffed animals can be given to

six children so that each child receives at least one but no more than three stuffed animals.

16. Use generating functions to find the number of ways to choose a dozen bagels from three varieties—egg, salty, and plain—if at least two bagels of each kind but no more than three salty bagels are chosen.

17. In how many ways can 25 identical donuts be distributed to four police officers so that each officer gets at least three but no more than seven donuts?

18. Use generating functions to find the number of ways to select 14 balls from a jar containing 100 red balls, 100 blue balls, and 100 green balls so that no fewer than 3 and no more than 10 blue balls are selected. Assume that the order in which the balls are drawn does not matter.

19. What is the generating function for the sequence $\{c_k\}$, where c_k is the number of ways to make change for k dollars using \$1 bills, \$2 bills, \$5 bills, and \$10 bills?

20. What is the generating function for the sequence $\{c_k\}$, where c_k represents the number of ways to make change for k pesos using bills worth 10 pesos, 20 pesos, 50 pesos, and 100 pesos?

21. Give a combinatorial interpretation of the coefficient of x^4 in the expansion $(1 + x + x^2 + x^3 + \cdots)^3$. Use this interpretation to find this number.

22. Give a combinatorial interpretation of the coefficient of x^6 in the expansion $(1 + x + x^2 + x^3 + \cdots)^n$. Use this interpretation to find this number.

23. **a)** What is the generating function for $\{a_k\}$, where a_k is the number of solutions of $x_1 + x_2 + x_3 = k$ when x_1, x_2, and x_3 are integers with $x_1 \geq 2$, $0 \leq x_2 \leq 3$, and $2 \leq x_3 \leq 5$?

b) Use your answer to part (a) to find a_6.

24. **a)** What is the generating function for $\{a_k\}$, where a_k is the number of solutions of $x_1 + x_2 + x_3 + x_4 = k$ when x_1, x_2, x_3, and x_4 are integers with $x_1 \geq 3$, $1 \leq x_2 \leq 5$, $0 \leq x_3 \leq 4$, and $x_4 \geq 1$?

b) Use your answer to part (a) to find a_7.

25. Explain how generating functions can be used to find the number of ways in which postage of r cents can be pasted on an envelope using 3-cent, 4-cent, and 20-cent stamps.

a) Assume that the order the stamps are pasted on does not matter.

b) Assume that the stamps are pasted in a row and the order in which they are pasted on matters.

c) Use your answer to part (a) to determine the number of ways 46 cents of postage can be pasted on an envelope using 3-cent, 4-cent, and 20-cent stamps when the order the stamps are pasted on does not matter. (Use of a computer algebra program is advised.)

d) Use your answer to part (b) to determine the number of ways 46 cents of postage can be pasted in a row on an envelope using 3-cent, 4-cent, and 20-cent stamps when the order in which the stamps are pasted on matters. (Use of a computer algebra program is advised.)

26. **a)** Show that $1/(1 - x - x^2 - x^3 - x^4 - x^5 - x^6)$ is the generating function for the number of ways that the sum n can be obtained when a die is rolled repeatedly and the order of the rolls matters.

b) Use part (a) to find the number of ways to roll a total of 8 when a die is rolled repeatedly, and the order of the rolls matters. (Use of a computer algebra package is advised.)

27. Use generating functions (and a computer algebra package, if available) to find the number of ways to make change for \$1 using

a) dimes and quarters.

b) nickels, dimes, and quarters.

c) pennies, dimes, and quarters.

d) pennies, nickels, dimes, and quarters.

28. Use generating functions (and a computer algebra package, if available) to find the number of ways to make change for \$1 using pennies, nickels, dimes, and quarters with

a) no more than 10 pennies.

b) no more than 10 pennies and no more than 10 nickels.

***c)** no more than 10 coins.

29. Use generating functions to find the number of ways to make change for \$100 using

a) \$10, \$20, and \$50 bills.

b) \$5, \$10, \$20, and \$50 bills.

c) \$5, \$10, \$20, and \$50 bills if at least one bill of each denomination is used.

d) \$5, \$10, and \$20 bills if at least one and no more than four of each denomination is used.

30. If $G(x)$ is the generating function for the sequence $\{a_k\}$, what is the generating function for each of these sequences?

a) $2a_0, 2a_1, 2a_2, 2a_3, \ldots$

b) $0, a_0, a_1, a_2, a_3, \ldots$ (assuming that terms follow the pattern of all but the first term)

c) $0, 0, 0, 0, a_2, a_3, \ldots$ (assuming that terms follow the pattern of all but the first four terms)

d) $a_2, a_3, a_4, \ldots$

e) $a_1, 2a_2, 3a_3, 4a_4, \ldots$ [*Hint: Calculus required* here.]

f) $a_0^2,\ 2a_0a_1,\ a_1^2 + 2a_0a_2,\ 2a_0a_3 + 2a_1a_2,\ 2a_0a_4 + 2a_1a_3 + a_2^2, \ldots$

31. If $G(x)$ is the generating function for the sequence $\{a_k\}$, what is the generating function for each of these sequences?

a) $0, 0, 0, a_3, a_4, a_5, \ldots$ (assuming that terms follow the pattern of all but the first three terms)

b) $a_0, 0, a_1, 0, a_2, 0, \ldots$

c) $0, 0, 0, 0, a_0, a_1, a_2, \ldots$ (assuming that terms follow the pattern of all but the first four terms)

d) $a_0, 2a_1, 4a_2, 8a_3, 16a_4, \ldots$

e) $0, a_0, a_1/2, a_2/3, a_3/4, \ldots$ [*Hint: Calculus required* here.]

f) $a_0, a_0 + a_1, a_0 + a_1 + a_2, a_0 + a_1 + a_2 + a_3, \ldots$

32. Use generating functions to solve the recurrence relation $a_k = 7a_{k-1}$ with the initial condition $a_0 = 5$.

33. Use generating functions to solve the recurrence relation $a_k = 3a_{k-1} + 2$ with the initial condition $a_0 = 1$.

34. Use generating functions to solve the recurrence relation $a_k = 3a_{k-1} + 4^{k-1}$ with the initial condition $a_0 = 1$.

35. Use generating functions to solve the recurrence relation $a_k = 5a_{k-1} - 6a_{k-2}$ with initial conditions $a_0 = 6$ and $a_1 = 30$.

36. Use generating functions to solve the recurrence relation $a_k = a_{k-1} + 2a_{k-2} + 2^k$ with initial conditions $a_0 = 4$ and $a_1 = 12$.

37. Use generating functions to solve the recurrence relation $a_k = 4a_{k-1} - 4a_{k-2} + k^2$ with initial conditions $a_0 = 2$ and $a_1 = 5$.

38. Use generating functions to solve the recurrence relation $a_k = 2a_{k-1} + 3a_{k-2} + 4^k + 6$ with initial conditions $a_0 = 20$, $a_1 = 60$.

39. Use generating functions to find an explicit formula for the Fibonacci numbers.

***40. a)** Show that if n is a positive integer, then

$$\binom{-1/2}{n} = \frac{\binom{2n}{n}}{(-4)^n}.$$

b) Use the extended Binomial Theorem and part (a) to show that the coefficient of x^n in the expansion of $(1-4x)^{-1/2}$ is $\binom{2n}{n}$ for all nonnegative integers n.

***41.** (*Calculus required*) Let $\{C_n\}$ be the sequence of Catalan numbers, that is, the solution to the recurrence relation $C_n = \sum_{k=0}^{n-1} C_k C_{n-k-1}$ with $C_0 = C_1 = 1$ (see Example 8 in Section 7.1).

a) Show that if $G(x)$ is the generating function for the sequence of Catalan numbers, then $xG(x)^2 - G(x) + 1 = 0$. Conclude (using the initial conditions) that $G(x) = (1 - \sqrt{1-4x})/(2x)$.

b) Use Exercise 40 to conclude that

$$G(x) = \sum_{n=0}^{\infty} \frac{1}{n+1}\binom{2n}{n} x^n,$$

so that

$$C_n = \frac{1}{n+1}\binom{2n}{n}.$$

42. Use generating functions to prove Pascal's identity: $C(n,r) = C(n-1,r) + C(n-1,r-1)$ when n and r are positive integers with $r < n$. [*Hint:* Use the identity $(1+x)^n = (1+x)^{n-1} + x(1+x)^{n-1}$.]

43. Use generating functions to prove Vandermonde's identity: $C(m+n,r) = \sum_{k=0}^{r} C(m,r-k)C(n,k)$, whenever m, n, and r are nonnegative integers with r not exceeding either m or n. [*Hint:* Look at the coefficient of x^r in both sides of $(1+x)^{m+n} = (1+x)^m(1+x)^n$.]

44. This exercise shows how to use generating functions to derive a formula for the sum of the first n squares.

a) Show that $(x^2+x)/(1-x)^4$ is the generating function for the sequence $\{a_n\}$, where $a_n = 1^2 + 2^2 + \cdots + n^2$.

b) Use part (a) to find an explicit formula for the sum $1^2 + 2^2 + \cdots + n^2$.

The **exponential generating function** for the sequence $\{a_n\}$ is the series

$$\sum_{n=0}^{\infty} \frac{a_n}{n!} x^n.$$

For example, the exponential generating function for the sequence 1, 1, 1, ... is the function $\sum_{n=0}^{\infty} x^n/n! = e^x$. (You will find this particular series useful in these exercises.) Note that e^x is the (ordinary) generating function for the sequence 1, 1, 1/2!, 1/3!, 1/4!, ...

45. Find a closed form for the exponential generating function for the sequence $\{a_n\}$, where

a) $a_n = 2$. **b)** $a_n = (-1)^n$.
c) $a_n = 3^n$. **d)** $a_n = n+1$.
e) $a_n = 1/(n+1)$.

46. Find a closed form for the exponential generating function for the sequence $\{a_n\}$, where

a) $a_n = (-2)^n$. **b)** $a_n = -1$. **c)** $a_n = n$.
d) $a_n = n(n-1)$. **e)** $a_n = 1/((n+1)(n+2))$.

47. Find the sequence with each of these functions as its exponential generating function.

a) $f(x) = e^{-x}$ **b)** $f(x) = 3x^{2x}$
c) $f(x) = e^{3x} - 3e^{2x}$ **d)** $f(x) = (1-x) + e^{-2x}$
e) $f(x) = e^{-2x} - (1/(1-x))$
f) $f(x) = e^{-3x} - (1+x) + (1/(1-2x))$
g) $f(x) = e^{x^2}$

48. Find the sequence with each of these functions as its exponential generating function.

a) $f(x) = e^{3x}$ **b)** $f(x) = 2e^{-3x+1}$
c) $f(x) = e^{4x} + e^{-4x}$ **d)** $f(x) = (1+2x) + e^{3x}$
e) $f(x) = e^x - (1/(1+x))$
f) $f(x) = xe^x$ **g)** $f(x) = e^{x^3}$

49. A coding system encodes messages using strings of octal (base 8) digits. A codeword is considered valid if and only if it contains an even number of 7s.

a) Find a linear nonhomogeneous recurrence relation for the number of valid codewords of length n. What are the initial conditions?

b) Solve this recurrence relation using Theorem 6 in Section 7.2.

c) Solve this recurrence relation using generating functions.

***50.** A coding system encodes messages using strings of base 4 digits (that is, digits from the set {0, 1, 2, 3}). A codeword is valid if and only if it contains an even number of 0s and an even number of 1s. Let a_n equal the number of valid codewords of length n. Furthermore, let b_n, c_n, and d_n equal the number of strings of base 4 digits of length n with an even number of 0s and an odd number of 1s, with an odd number of 0s and an even number of

1s, and with an odd number of 0s and an odd number of 1s, respectively.

a) Show that $d_n = 4^n - a_n - b_n - c_n$. Use this to show that $a_{n+1} = 2a_n + b_n + c_n$, $b_{n+1} = b_n - c_n + 4^n$, and $c_{n+1} = c_n - b_n + 4^n$.
b) What are a_1, b_1, c_1, and d_1?
c) Use parts (a) and (b) to find a_3, b_3, c_3, and d_3.
d) Use the recurrence relations in part (a), together with the initial conditions in part (b), to set up three equations relating the generating functions $A(x)$, $B(x)$, and $C(x)$ for the sequences $\{a_n\}$, $\{b_n\}$, and $\{c_n\}$, respectively.
e) Solve the system of equations from part (d) to get explicit formulae for $A(x)$, $B(x)$, and $C(x)$ and use these to get explicit formulae for a_n, b_n, c_n, and d_n.

Generating functions are useful in studying the number of different types of partitions of an integer n. A **partition** of a positive integer is a way to write this integer as the sum of positive integers where repetition is allowed and the order of the integers in the sum does not matter. For example, the partitions of 5 (with no restrictions) are $1+1+1+1+1$, $1+1+1+2$, $1+1+3$, $1+2+2$, $1+4$, $2+3$, and 5. Exercises 51–56 illustrate some of these uses.

51. Show that the coefficient $p(n)$ of x^n in the formal power series expansion of $1/((1-x)(1-x^2)(1-x^3)\cdots)$ equals the number of partitions of n.

52. Show that the coefficient $p_o(n)$ of x^n in the formal power series expansion of $1/((1-x)(1-x^3)(1-x^5)\cdots)$ equals the number of partitions of n into odd integers, that is, the number of ways to write n as the sum of odd positive integers, where the order does not matter and repetitions are allowed.

53. Show that the coefficient $p_d(n)$ of x^n in the formal power series expansion of $(1+x)(1+x^2)(1+x^3)\cdots$ equals the number of partitions of n into distinct parts, that is, the number of ways to write n as the sum of positive integers, where the order does not matter but no repetitions are allowed.

54. Find $p_o(n)$, the number of partitions of n into odd parts with repetitions allowed, and $p_d(n)$, the number of partitions of n into distinct parts, for $1 \le n \le 8$, by writing each partition of each type for each integer.

55. Show that if n is a positive integer, then the number of partitions of n into distinct parts equals the number of partitions of n into odd parts with repetitions allowed; that is, $p_o(n) = p_d(n)$. [*Hint:* Show that the generating functions for $p_o(n)$ and $p_d(n)$ are equal.]

****56.** (*Requires calculus*) Use the generating function of $p(n)$ to show that $p(n) \le e^{C\sqrt{n}}$ for some constant C. [Hardy and Ramanujan showed that $p(n) \sim e^{\pi\sqrt{2/3}\sqrt{n}}/(4\sqrt{3}n)$, which means that the ratio of $p(n)$ and the right-hand side approaches 1 as n approaches infinity.]

Suppose that X is a random variable on a sample space S such that $X(s)$ is a nonnegative integer for all $s \in S$. The **probability generating function** for X is

$$G_X(x) = \sum_{k=0}^{\infty} p(X(s) = k)x^k.$$

57. (*Requires calculus*) Show that if G_X is the probability generating function for a random variable X such that $X(s)$ is a nonnegative integer for all $s \in S$, then

a) $G_X(1) = 1$.
b) $E(X) = G_X'(1)$.
c) $V(X) = G_X''(1) + G_X'(1) - G_X'(1)^2$.

58. Let X be the random variable whose value is n if the first success occurs on the nth trial when independent Bernoulli trials are performed, each with probability of success p.

a) Find a closed formula for the probability generating function G_X.
b) Find the expected value and the variance of X using Exercise 57 and the closed form for the probability generating function found in part (a).

59. Let m be a positive integer. Let X_m be the random variable whose value is n if the mth success occurs on the $(n+m)$th trial when independent Bernoulli trials are performed, each with probability of success p.

a) Using Exercise 28 in the Supplementary Exercises of Chapter 6, show that the probability generating function G_{X_m} is given by $G_{X_m}(x) = p^m/(1-qx)^m$, where $q = 1-p$.
b) Find the expected value and the variance of X_m using Exercise 57 and the closed form for the probability generating function in part (a).

60. Show that if X and Y are independent random variables on a sample space S such that $X(s)$ and $Y(s)$ are nonnegative integers for all $s \in S$, then $G_{X+Y}(x) = G_X(x)G_Y(x)$.

7.5 Inclusion–Exclusion

Introduction

A discrete mathematics class contains 30 women and 50 sophomores. How many students in the class are either women or sophomores? This question cannot be answered unless more information is provided. Adding the number of women in the class and the number of sophomores probably does not give the correct answer, because women sophomores are counted twice. This observation shows that the number of students in the class that are either sophomores or women

is the sum of the number of women and the number of sophomores in the class minus the number of women sophomores. A technique for solving such counting problems was introduced in Section 5.1. In this section we will generalize the ideas introduced in that section to solve problems that require us to count the number of elements in the union of more than two sets.

The Principle of Inclusion–Exclusion

How many elements are in the union of two finite sets? In Section 2.2 we showed that the number of elements in the union of the two sets A and B is the sum of the numbers of elements in the sets minus the number of elements in their intersection. That is,

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

As we showed in Section 5.1, the formula for the number of elements in the union of two sets is useful in counting problems. Examples 1–3 provide additional illustrations of the usefulness of this formula.

EXAMPLE 1 In a discrete mathematics class every student is a major in computer science or mathematics, or both. The number of students having computer science as a major (possibly along with mathematics) is 25; the number of students having mathematics as a major (possibly along with computer science) is 13; and the number of students majoring in both computer science and mathematics is 8. How many students are in this class?

Solution: Let A be the set of students in the class majoring in computer science and B be the set of students in the class majoring in mathematics. Then $A \cap B$ is the set of students in the class who are joint mathematics and computer science majors. Because every student in the class is majoring in either computer science or mathematics (or both), it follows that the number of students in the class is $|A \cup B|$. Therefore,

$$\begin{aligned} |A \cup B| &= |A| + |B| - |A \cap B| \\ &= 25 + 13 - 8 \\ &= 30. \end{aligned}$$

Therefore, there are 30 students in the class. This computation is illustrated in Figure 1. ◀

EXAMPLE 2 How many positive integers not exceeding 1000 are divisible by 7 or 11?

Solution: Let A be the set of positive integers not exceeding 1000 that are divisible by 7, and let B be the set of positive integers not exceeding 1000 that are divisible by 11. Then $A \cup B$ is the set of integers not exceeding 1000 that are divisible by either 7 or 11, and $A \cap B$ is the set of integers not exceeding 1000 that are divisible by both 7 and 11. From Example 2 of Section 3.4, we know that among the positive integers not exceeding 1000 there are $\lfloor 1000/7 \rfloor$ integers divisible by 7 and $\lfloor 1000/11 \rfloor$ divisible by 11. Because 7 and 11 are relatively prime, the integers divisible by both 7 and 11 are those divisible by $7 \cdot 11$. Consequently, there are $\lfloor 1000/(11 \cdot 7) \rfloor$ positive integers not exceeding 1000 that are divisible by both 7 and 11. It follows that there are

$$\begin{aligned} |A \cup B| &= |A| + |B| - |A \cap B| \\ &= \left\lfloor \frac{1000}{7} \right\rfloor + \left\lfloor \frac{1000}{11} \right\rfloor - \left\lfloor \frac{1000}{7 \cdot 11} \right\rfloor \\ &= 142 + 90 - 12 \\ &= 220 \end{aligned}$$

FIGURE 1 The Set of Students in a Discrete Mathematics Class.

FIGURE 2 The Set of Positive Integers Not Exceeding 1000 Divisible by Either 7 or 11.

positive integers not exceeding 1000 that are divisible by either 7 or 11. This computation is illustrated in Figure 2. ◀

Example 3 shows how to find the number of elements in a finite universal set that are outside the union of two sets.

EXAMPLE 3 Suppose that there are 1807 freshmen at your school. Of these, 453 are taking a course in computer science, 567 are taking a course in mathematics, and 299 are taking courses in both computer science and mathematics. How many are not taking a course either in computer science or in mathematics?

Solution: To find the number of freshmen who are not taking a course in either mathematics or computer science, subtract the number that are taking a course in either of these subjects from the total number of freshmen. Let A be the set of all freshmen taking a course in computer science, and let B be the set of all freshmen taking a course in mathematics. It follows that $|A| = 453$, $|B| = 567$, and $|A \cap B| = 299$. The number of freshmen taking a course in either computer science or mathematics is

$$|A \cup B| = |A| + |B| - |A \cap B| = 453 + 567 - 299 = 721.$$

Consequently, there are $1807 - 721 = 1086$ freshmen who are not taking a course in computer science or mathematics. ◀

We will now begin our development of a formula for the number of elements in the union of a finite number of sets. The formula we will develop is called the **principle of inclusion–exclusion.** For concreteness, before we consider unions of n sets, where n is any positive integer, we will derive a formula for the number of elements in the union of three sets A, B, and C. To construct this formula, we note that $|A| + |B| + |C|$ counts each element that is in exactly one of the three sets once, elements that are in exactly two of the sets twice, and elements in all three sets three times. This is illustrated in the first panel in Figure 3.

To remove the overcount of elements in more than one of the sets, we subtract the number of elements in the intersections of all pairs of the three sets. We obtain

$$|A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C|.$$

(a) Count of elements by $|A|+|B|+|C|$

(b) Count of elements by $|A|+|B|+|C|-|A\cap B|-|A\cap C|-|B\cap C|$

(c) Count of elements by $|A|+|B|+|C|-|A\cap B|-|A\cap C|-|B\cap C|+|A\cap B\cap C|$

FIGURE 3 Finding a Formula for the Number of Elements in the Union of Three Sets.

This expression still counts elements that occur in exactly one of the sets once. An element that occurs in exactly two of the sets is also counted exactly once, because this element will occur in one of the three intersections of sets taken two at a time. However, those elements that occur in all three sets will be counted zero times by this expression, because they occur in all three intersections of sets taken two at a time. This is illustrated in the second panel in Figure 3.

To remedy this undercount, we add the number of elements in the intersection of all three sets. This final expression counts each element once, whether it is in one, two, or three of the sets. Thus,

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

This formula is illustrated in the third panel of Figure 3.

Example 4 illustrates how this formula can be used.

EXAMPLE 4 A total of 1232 students have taken a course in Spanish, 879 have taken a course in French, and 114 have taken a course in Russian. Further, 103 have taken courses in both Spanish and French, 23 have taken courses in both Spanish and Russian, and 14 have taken courses in both French and Russian. If 2092 students have taken at least one of Spanish, French, and Russian, how many students have taken a course in all three languages?

Solution: Let S be the set of students who have taken a course in Spanish, F the set of students who have taken a course in French, and R the set of students who have taken a course in Russian. Then

$$|S| = 1232, \quad |F| = 879, \quad |R| = 114,$$
$$|S \cap F| = 103, \; |S \cap R| = 23, \; |F \cap R| = 14,$$

and

$$|S \cup F \cup R| = 2092.$$

When we insert these quantities into the equation

$$|S \cup F \cup R| = |S| + |F| + |R| - |S \cap F| - |S \cap R| - |F \cap R| + |S \cap F \cap R|$$

FIGURE 4 The Set of Students Who Have Taken Courses in Spanish, French, and Russian.

we obtain

$$2092 = 1232 + 879 + 114 - 103 - 23 - 14 + |S \cap F \cap R|.$$

We now solve for $|S \cap F \cap R|$. We find that $|S \cap F \cap R| = 7$. Therefore, there are seven students who have taken courses in Spanish, French, and Russian. This is illustrated in Figure 4. ◀

We will now state and prove the inclusion–exclusion principle, which tells us how many elements are in the union of a finite number of finite sets.

THEOREM 1 **THE PRINCIPLE OF INCLUSION–EXCLUSION** Let $A_1, A_2, \ldots, A_n$ be finite sets. Then

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = \sum_{1 \le i \le n} |A_i| - \sum_{1 \le i < j \le n} |A_i \cap A_j| + \sum_{1 \le i < j < k \le n} |A_i \cap A_j \cap A_k| - \cdots + (-1)^{n+1} |A_1 \cap A_2 \cap \cdots \cap A_n|.$$

Proof: We will prove the formula by showing that an element in the union is counted exactly once by the right-hand side of the equation. Suppose that a is a member of exactly r of the sets $A_1, A_2, \ldots, A_n$ where $1 \le r \le n$. This element is counted $C(r, 1)$ times by $\Sigma|A_i|$. It is counted $C(r, 2)$ times by $\Sigma|A_i \cap A_j|$. In general, it is counted $C(r, m)$ times by the summation involving m of the sets A_i. Thus, this element is counted exactly

$$C(r, 1) - C(r, 2) + C(r, 3) - \cdots + (-1)^{r+1} C(r, r)$$

times by the expression on the right-hand side of this equation. Our goal is to evaluate this quantity. By Corollary 2 of Section 5.4, we have

$$C(r, 0) - C(r, 1) + C(r, 2) - \cdots + (-1)^r C(r, r) = 0.$$

Hence,

$$1 = C(r, 0) = C(r, 1) - C(r, 2) + \cdots + (-1)^{r+1} C(r, r).$$

Therefore, each element in the union is counted exactly once by the expression on the right-hand side of the equation. This proves the principle of inclusion–exclusion. ◁

The inclusion–exclusion principle gives a formula for the number of elements in the union of n sets for every positive integer n. There are terms in this formula for the number of elements in the intersection of every nonempty subset of the collection of the n sets. Hence, there are $2^n - 1$ terms in this formula.

EXAMPLE 5 Give a formula for the number of elements in the union of four sets.

Extra Examples

Solution: The inclusion–exclusion principle shows that

$$\begin{aligned}|A_1 \cup A_2 \cup A_3 \cup A_4| = {} & |A_1| + |A_2| + |A_3| + |A_4| \\ & - |A_1 \cap A_2| - |A_1 \cap A_3| - |A_1 \cap A_4| - |A_2 \cap A_3| - |A_2 \cap A_4| \\ & - |A_3 \cap A_4| + |A_1 \cap A_2 \cap A_3| + |A_1 \cap A_2 \cap A_4| + |A_1 \cap A_3 \cap A_4| \\ & + |A_2 \cap A_3 \cap A_4| - |A_1 \cap A_2 \cap A_3 \cap A_4|.\end{aligned}$$

Note that this formula contains 15 different terms, one for each nonempty subset of $\{A_1, A_2, A_3, A_4\}$. ◀

Exercises

1. How many elements are in $A_1 \cup A_2$ if there are 12 elements in A_1, 18 elements in A_2, and
   a) $A_1 \cap A_2 = \emptyset$? **b)** $|A_1 \cap A_2| = 1$?
   c) $|A_1 \cap A_2| = 6$? **d)** $A_1 \subseteq A_2$?
2. There are 345 students at a college who have taken a course in calculus, 212 who have taken a course in discrete mathematics, and 188 who have taken courses in both calculus and discrete mathematics. How many students have taken a course in either calculus or discrete mathematics?
3. A survey of households in the United States reveals that 96% have at least one television set, 98% have telephone service, and 95% have telephone service and at least one television set. What percentage of households in the United States have neither telephone service nor a television set?
4. A marketing report concerning personal computers states that 650,000 owners will buy a modem for their machines next year and 1,250,000 will buy at least one software package. If the report states that 1,450,000 owners will buy either a modem or at least one software package, how many will buy both a modem and at least one software package?
5. Find the number of elements in $A_1 \cup A_2 \cup A_3$ if there are 100 elements in each set and if
   a) the sets are pairwise disjoint.
   b) there are 50 common elements in each pair of sets and no elements in all three sets.
   c) there are 50 common elements in each pair of sets and 25 elements in all three sets.
   d) the sets are equal.
6. Find the number of elements in $A_1 \cup A_2 \cup A_3$ if there are 100 elements in A_1, 1000 in A_2, and 10,000 in A_3 if
   a) $A_1 \subseteq A_2$ and $A_2 \subseteq A_3$.
   b) the sets are pairwise disjoint.
   c) there are two elements common to each pair of sets and one element in all three sets.
7. There are 2504 computer science students at a school. Of these, 1876 have taken a course in Pascal, 999 have taken a course in Fortran, and 345 have taken a course in C.

Further, 876 have taken courses in both Pascal and Fortran, 231 have taken courses in both Fortran and C, and 290 have taken courses in both Pascal and C. If 189 of these students have taken courses in Fortran, Pascal, and C, how many of these 2504 students have not taken a course in any of these three programming languages?

8. In a survey of 270 college students, it is found that 64 like brussels sprouts, 94 like broccoli, 58 like cauliflower, 26 like both brussels sprouts and broccoli, 28 like both brussels sprouts and cauliflower, 22 like both broccoli and cauliflower, and 14 like all three vegetables. How many of the 270 students do not like any of these vegetables?

9. How many students are enrolled in a course either in calculus, discrete mathematics, data structures, or programming languages at a school if there are 507, 292, 312, and 344 students in these courses, respectively; 14 in both calculus and data structures; 213 in both calculus and programming languages; 211 in both discrete mathematics and data structures; 43 in both discrete mathematics and programming languages; and no student may take calculus and discrete mathematics, or data structures and programming languages, concurrently?

10. Find the number of positive integers not exceeding 100 that are not divisible by 5 or by 7.

11. Find the number of positive integers not exceeding 100 that are either odd or the square of an integer.

12. Find the number of positive integers not exceeding 1000 that are either the square or the cube of an integer.

13. How many bit strings of length eight do not contain six consecutive 0s?

***14.** How many permutations of the 26 letters of the English alphabet do not contain any of the strings *fish, rat,* or *bird*?

15. How many permutations of the 10 digits either begin with the 3 digits 987, contain the digits 45 in the fifth and sixth positions, or end with the 3 digits 123?

16. How many elements are in the union of four sets if each of the sets has 100 elements, each pair of the sets shares 50 elements, each three of the sets share 25 elements, and there are 5 elements in all four sets?

17. How many elements are in the union of four sets if the sets have 50, 60, 70, and 80 elements, respectively, each pair of the sets has 5 elements in common, each triple of the sets has 1 common element, and no element is in all four sets?

18. How many terms are there in the formula for the number of elements in the union of 10 sets given by the principle of inclusion–exclusion?

19. Write out the explicit formula given by the principle of inclusion–exclusion for the number of elements in the union of five sets.

20. How many elements are in the union of five sets if the sets contain 10,000 elements each, each pair of sets has 1000 common elements, each triple of sets has 100 common elements, every four of the sets have 10 common elements, and there is 1 element in all five sets?

21. Write out the explicit formula given by the principle of inclusion–exclusion for the number of elements in the union of six sets when it is known that no three of these sets have a common intersection.

***22.** Prove the principle of inclusion–exclusion using mathematical induction.

23. Let E_1, E_2, and E_3 be three events from a sample space S. Find a formula for the probability of $E_1 \cup E_2 \cup E_3$.

24. Find the probability that when a fair coin is flipped five times tails comes up exactly three times, the first and last flips come up tails, or the second and fourth flips come up heads.

25. Find the probability that when four numbers from 1 to 100, inclusive, are picked at random with no repetitions allowed, either all are odd, all are divisible by 3, or all are divisible by 5.

26. Find a formula for the probability of the union of four events in a sample space if no three of them can occur at the same time.

27. Find a formula for the probability of the union of five events in a sample space if no four of them can occur at the same time.

28. Find a formula for the probability of the union of n events in a sample space when no two of these events can occur at the same time.

29. Find a formula for the probability of the union of n events in a sample space.

7.6 Applications of Inclusion–Exclusion

Introduction

Many counting problems can be solved using the principle of inclusion–exclusion. For instance, we can use this principle to find the number of primes less than a positive integer. Many problems can be solved by counting the number of onto functions from one finite set to another. The inclusion–exclusion principle can be used to find the number of such functions. The famous hatcheck problem can be solved using the principle of inclusion–exclusion. This problem asks for the probability that no person is given the correct hat back by a hatcheck person who gives the hats back randomly.

An Alternative Form of Inclusion–Exclusion

There is an alternative form of the principle of inclusion–exclusion that is useful in counting problems. In particular, this form can be used to solve problems that ask for the number of elements in a set that have none of n properties $P_1, P_2, \ldots, P_n$.

Let A_i be the subset containing the elements that have property P_i. The number of elements with all the properties $P_{i_1}, P_{i_2}, \ldots, P_{i_k}$ will be denoted by $N(P_{i_1}P_{i_2}\cdots P_{i_k})$. Writing these quantities in terms of sets, we have

$$|A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}| = N(P_{i_1}P_{i_2}\cdots P_{i_k}).$$

If the number of elements with none of the properties $P_1, P_2, \ldots, P_n$ is denoted by $N(P_1'P_2'\cdots P_n')$ and the number of elements in the set is denoted by N, it follows that

$$N(P_1'P_2'\cdots P_n') = N - |A_1 \cup A_2 \cup \cdots \cup A_n|.$$

From the inclusion–exclusion principle, we see that

$$\begin{aligned} N(P_1'P_2'\cdots P_n') = N &- \sum_{1\le i\le n} N(P_i) + \sum_{1\le i<j\le n} N(P_iP_j) \\ &- \sum_{1\le i<j<k\le n} N(P_iP_jP_k) + \cdots + (-1)^n N(P_1P_2\cdots P_n). \end{aligned}$$

Example 1 shows how the principle of inclusion–exclusion can be used to determine the number of solutions in integers of an equation with constraints.

EXAMPLE 1 How many solutions does

$$x_1 + x_2 + x_3 = 11$$

have, where x_1, x_2, and x_3 are nonnegative integers with $x_1 \le 3$, $x_2 \le 4$, and $x_3 \le 6$?

Solution: To apply the principle of inclusion–exclusion, let a solution have property P_1 if $x_1 > 3$, property P_2 if $x_2 > 4$, and property P_3 if $x_3 > 6$. The number of solutions satisfying the inequalities $x_1 \le 3$, $x_2 \le 4$, and $x_3 \le 6$ is

$$\begin{aligned} N(P_1'P_2'P_3') = N - N(P_1) - N(P_2) - N(P_3) + N(P_1P_2) \\ + N(P_1P_3) + N(P_2P_3) - N(P_1P_2P_3). \end{aligned}$$

Using the same techniques as in Example 5 of Section 5.5, it follows that

- N = total number of solutions $= C(3 + 11 - 1, 11) = 78$,
- $N(P_1)$ = (number of solutions with $x_1 \ge 4$) $= C(3 + 7 - 1, 7) = C(9, 7) = 36$,
- $N(P_2)$ = (number of solutions with $x_2 \ge 5$) $= C(3 + 6 - 1, 6) = C(8, 6) = 28$,
- $N(P_3)$ = (number of solutions with $x_3 \ge 7$) $= C(3 + 4 - 1, 4) = C(6, 4) = 15$,
- $N(P_1P_2)$ = (number of solutions with $x_1 \ge 4$ and $x_2 \ge 5$) $= C(3 + 2 - 1, 2) = C(4, 2) = 6$,
- $N(P_1P_3)$ = (number of solutions with $x_1 \ge 4$ and $x_3 \ge 7$) $= C(3 + 0 - 1, 0) = 1$,
- $N(P_2P_3)$ = (number of solutions with $x_2 \ge 5$ and $x_3 \ge 7$) $= 0$,
- $N(P_1P_2P_3)$ = (number of solutions with $x_1 \ge 4$, $x_2 \ge 5$, and $x_3 \ge 7$) $= 0$.

Inserting these quantities into the formula for $N(P_1'P_2'P_3')$ shows that the number of solutions with $x_1 \le 3$, $x_2 \le 4$, and $x_3 \le 6$ equals

$$N(P_1'P_2'P_3') = 78 - 36 - 28 - 15 + 6 + 1 + 0 - 0 = 6.$$ ◀

The Sieve of Eratosthenes

The principle of inclusion–exclusion can be used to find the number of primes not exceeding a specified positive integer. Recall that a composite integer is divisible by a prime not exceeding its square root. So, to find the number of primes not exceeding 100, first note that composite integers not exceeding 100 must have a prime factor not exceeding 10. Because the only primes less than 10 are 2, 3, 5, and 7, the primes not exceeding 100 are these four primes and those positive integers greater than 1 and not exceeding 100 that are divisible by none of 2, 3, 5, or 7. To apply the principle of inclusion–exclusion, let P_1 be the property that an integer is divisible by 2, let P_2 be the property that an integer is divisible by 3, let P_3 be the property that an integer is divisible by 5, and let P_4 be the property that an integer is divisible by 7. Thus, the number of primes not exceeding 100 is

$$4 + N(P_1'P_2'P_3'P_4').$$

Because there are 99 positive integers greater than 1 and not exceeding 100, the principle of inclusion–exclusion shows that

$$\begin{aligned} N(P_1'P_2'P_3'P_4') = {} & 99 - N(P_1) - N(P_2) - N(P_3) - N(P_4) \\ & + N(P_1P_2) + N(P_1P_3) + N(P_1P_4) + N(P_2P_3) + N(P_2P_4) + N(P_3P_4) \\ & - N(P_1P_2P_3) - N(P_1P_2P_4) - N(P_1P_3P_4) - N(P_2P_3P_4) \\ & + N(P_1P_2P_3P_4). \end{aligned}$$

The number of integers not exceeding 100 (and greater than 1) that are divisible by all the primes in a subset of $\{2, 3, 5, 7\}$ is $\lfloor 100/N \rfloor$, where N is the product of the primes in this subset. (This follows because any two of these primes have no common factor.) Consequently,

$$\begin{aligned} N(P_1'P_2'P_3'P_4') = {} & 99 - \left\lfloor \frac{100}{2} \right\rfloor - \left\lfloor \frac{100}{3} \right\rfloor - \left\lfloor \frac{100}{5} \right\rfloor - \left\lfloor \frac{100}{7} \right\rfloor \\ & + \left\lfloor \frac{100}{2 \cdot 3} \right\rfloor + \left\lfloor \frac{100}{2 \cdot 5} \right\rfloor + \left\lfloor \frac{100}{2 \cdot 7} \right\rfloor + \left\lfloor \frac{100}{3 \cdot 5} \right\rfloor + \left\lfloor \frac{100}{3 \cdot 7} \right\rfloor + \left\lfloor \frac{100}{5 \cdot 7} \right\rfloor \\ & - \left\lfloor \frac{100}{2 \cdot 3 \cdot 5} \right\rfloor - \left\lfloor \frac{100}{2 \cdot 3 \cdot 7} \right\rfloor - \left\lfloor \frac{100}{2 \cdot 5 \cdot 7} \right\rfloor - \left\lfloor \frac{100}{3 \cdot 5 \cdot 7} \right\rfloor + \left\lfloor \frac{100}{2 \cdot 3 \cdot 5 \cdot 7} \right\rfloor \\ = {} & 99 - 50 - 33 - 20 - 14 + 16 + 10 + 7 + 6 + 4 + 2 - 3 - 2 - 1 - 0 + 0 \\ = {} & 21. \end{aligned}$$

Hence, there are $4 + 21 = 25$ primes not exceeding 100.

Lin!ts The **Sieve of Eratosthenes** is used to find all primes not exceeding a specified positive integer. For instance, the following procedure is used to find the primes not exceeding 100. First the integers that are divisible by 2, other than 2, are deleted. Because 3 is the first integer greater

TABLE 1 The Sieve of Eratosthenes.

Integers divisible by 2 other than 2 receive an underline.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Integers divisible by 3 other than 3 receive an underline.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Integers divisible by 5 other than 5 receive an underline.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Integers divisible by 7 other than 7 receive an underline; integers in color are prime.

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

than 2 that is left, all those integers divisible by 3, other than 3, are deleted. Because 5 is the next integer left after 3, those integers divisible by 5, other than 5, are deleted. The next integer left is 7, so those integers divisible by 7, other than 7, are deleted. Because all composite integers not exceeding 100 are divisible by 2, 3, 5, or 7, all remaining integers except 1 are prime. In Table 1, the panels display those integers deleted at each stage, where each integer divisible by 2, other than 2, is underlined in the first panel, each integer divisible by 3, other than 3, is underlined in the second panel, each integer divisible by 5, other than 5, is underlined in the third panel,

Links

ERATOSTHENES (276–194 B.C.E.) It is known that Eratosthenes was born in Cyrene, a Greek colony west of Egypt, and spent time studying at Plato's Academy in Athens. We also know that King Ptolemy II invited Eratosthenes to Alexandria to tutor his son and that later Eratosthenes became chief librarian at the famous library at Alexandria, a central repository of ancient wisdom. Eratosthenes was an extremely versatile scholar, writing on mathematics, geography, astronomy, history, philosophy, and literary criticism. Besides his work in mathematics, he is most noted for his chronology of ancient history and for his famous measurement of the size of the earth.

and each integer divisible by 7, other than 7, is underlined in the fourth panel. The integers not underlined are the primes not exceeding 100.

The Number of Onto Functions

The principle of inclusion–exclusion can also be used to determine the number of onto functions from a set with m elements to a set with n elements. First consider Example 2.

EXAMPLE 2 How many onto functions are there from a set with six elements to a set with three elements?

Solution: Suppose that the elements in the codomain are b_1, b_2, and b_3. Let P_1, P_2, and P_3 be the properties that b_1, b_2, and b_3 are not in the range of the function, respectively. Note that a function is onto if and only if it has none of the properties P_1, P_2, or P_3. By the inclusion–exclusion principle it follows that the number of onto functions from a set with six elements to a set with three elements is

$$N(P_1'P_2'P_3') = N - [N(P_1) + N(P_2) + N(P_3)] + [N(P_1P_2) + N(P_1P_3) + N(P_2P_3)] - N(P_1P_2P_3),$$

where N is the total number of functions from a set with six elements to one with three elements. We will evaluate each of the terms on the right-hand side of this equation.

From Example 6 of Section 5.1, it follows that $N = 3^6$. Note that $N(P_i)$ is the number of functions that do not have b_i in their range. Hence, there are two choices for the value of the function at each element of the domain. Therefore, $N(P_i) = 2^6$. Furthermore, there are $C(3, 1)$ terms of this kind. Note that $N(P_iP_j)$ is the number of functions that do not have b_i and b_j in their range. Hence, there is only one choice for the value of the function at each element of the domain. Therefore, $N(P_iP_j) = 1^6 = 1$. Furthermore, there are $C(3, 2)$ terms of this kind. Also, note that $N(P_1P_2P_3) = 0$, because this term is the number of functions that have none of b_1, b_2, and b_3 in their range. Clearly, there are no such functions. Therefore, the number of onto functions from a set with six elements to one with three elements is

$$3^6 - C(3, 1)2^6 + C(3, 2)1^6 = 729 - 192 + 3 = 540.$$ ◀

The general result that tells us how many onto functions there are from a set with m elements to one with n elements will now be stated. The proof of this result is left as an exercise for the reader.

THEOREM 1 Let m and n be positive integers with $m \geq n$. Then, there are

$$n^m - C(n, 1)(n-1)^m + C(n, 2)(n-2)^m - \cdots + (-1)^{n-1}C(n, n-1) \cdot 1^m$$

onto functions from a set with m elements to a set with n elements.

An onto function from a set with m elements to a set with n elements corresponds to a way to distribute the m elements in the domain to n indistinguishable boxes so that no box is empty, and then to associate each of the n elements of the codomain to a box. This means that the number of onto functions from a set with m elements to a set with n elements is the number of ways to distribute m distinguishable objects to n indistinguishable boxes so that no box is empty multiplied by the number of permutations of a set with n elements. Consequently, the number of onto functions from a set with m elements to a set with n elements equals $n!S(m, n)$,

where $S(m, n)$ is a *Stirling number of the second kind* defined in Section 5.5. This means that we can use Theorem 1 to deduce the formula given in Section 5.5 for $S(m, n)$. (See Chapter 6 of [MiRo91] for more details about Stirling numbers of the second kind.)

One of the many different applications of Theorem 1 will now be described.

EXAMPLE 3 How many ways are there to assign five different jobs to four different employees if every employee is assigned at least one job?

Solution: Consider the assignment of jobs as a function from the set of five jobs to the set of four employees. An assignment where every employee gets at least one job is the same as an onto function from the set of jobs to the set of employees. Hence, by Theorem 1 it follows that there are

$$4^5 - C(4, 1)3^5 + C(4, 2)2^5 - C(4, 3)1^5 = 1024 - 972 + 192 - 4 = 240$$

ways to assign the jobs so that each employee is assigned at least one job. ◀

Derangements

The principle of inclusion–exclusion will be used to count the permutations of n objects that leave no objects in their original positions. Consider Example 4.

EXAMPLE 4 **The Hatcheck Problem** A new employee checks the hats of n people at a restaurant, forgetting to put claim check numbers on the hats. When customers return for their hats, the checker gives them back hats chosen at random from the remaining hats. What is the probability that no one receives the correct hat? ◀

Remark: The answer is the number of ways the hats can be arranged so that there is no hat in its original position divided by $n!$, the number of permutations of n hats. We will return to this example after we find the number of permutations of n objects that leave no objects in their original position.

Links

A **derangement** is a permutation of objects that leaves no object in its original position. To solve the problem posed in Example 4 we will need to determine the number of derangements of a set of n objects.

EXAMPLE 5 The permutation 21453 is a derangement of 12345 because no number is left in its original position. However, 21543 is not a derangement of 12345, because this permutation leaves 4 fixed. ◀

Let D_n denote the number of derangements of n objects. For instance, $D_3 = 2$, because the derangements of 123 are 231 and 312. We will evaluate D_n, for all positive integers n, using the principle of inclusion–exclusion.

THEOREM 2 The number of derangements of a set with n elements is

$$D_n = n!\left[1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!}\right].$$

Proof: Let a permutation have property P_i if it fixes element i. The number of derangements is the number of permutations having none of the properties P_i for $i = 1, 2, \ldots, n$. This means that

$$D_n = N(P_1' P_2' \cdots P_n').$$

Using the principle of inclusion–exclusion, it follows that

$$D_n = N - \sum_i N(P_i) + \sum_{i<j} N(P_i P_j) - \sum_{i<j<k} N(P_i P_j P_k) + \cdots + (-1)^n N(P_1 P_2 \cdots P_n),$$

where N is the number of permutations of n elements. This equation states that the number of permutations that fix no elements equals the total number of permutations, less the number that fix at least one element, plus the number that fix at least two elements, less the number that fix at least three elements, and so on. All the quantities that occur on the right-hand side of this equation will now be found.

First, note that $N = n!$, because N is simply the total number of permutations of n elements. Also, $N(P_i) = (n-1)!$. This follows from the product rule, because $N(P_i)$ is the number of permutations that fix element i, so the ith position of the permutation is determined, but each of the remaining positions can be filled arbitrarily. Similarly,

$$N(P_i P_j) = (n-2)!,$$

because this is the number of permutations that fix elements i and j, but where the other $n-2$ elements can be arranged arbitrarily. In general, note that

$$N(P_{i_1} P_{i_2} \cdots P_{i_m}) = (n-m)!,$$

because this is the number of permutations that fix elements $i_1, i_2, \ldots, i_m$, but where the other $n-m$ elements can be arranged arbitrarily. Because there are $C(n, m)$ ways to choose m elements from n, it follows that

$$\sum_{1 \le i \le n} N(P_i) = C(n, 1)(n-1)!,$$

$$\sum_{1 \le i < j \le n} N(P_i P_j) = C(n, 2)(n-2)!,$$

and in general,

$$\sum_{1 \le i_1 < i_2 < \cdots < i_m \le n} N(P_{i_1} P_{i_2} \cdots P_{i_m}) = C(n, m)(n-m)!.$$

Links

HISTORICAL NOTE In *rencontres* (matches), an old French card game, the 52 cards in a deck are laid out in a row. The cards of a second deck are laid out with one card of the second deck on top of each card of the first deck. The score is determined by counting the number of matching cards in the two decks. In 1708 Pierre Raymond de Montmort (1678–1719) posed *le problème de rencontres:* What is the probability that no matches take place in the game of rencontres? The solution to Montmort's problem is the probability that a randomly selected permutation of 52 objects is a derangement, namely, $D_{52}/52!$, which, as we will see, is approximately $1/e$.

TABLE 2 The Probability of a Derangement.

n	2	3	4	5	6	7
$D_n/n!$	0.50000	0.33333	0.37500	0.36667	0.36806	0.36786

Consequently, inserting these quantities into our formula for D_n gives

$$\begin{aligned} D_n &= n! - C(n,1)(n-1)! + C(n,2)(n-2)! - \cdots + (-1)^n C(n,n)(n-n)! \\ &= n! - \frac{n!}{1!(n-1)!}(n-1)! + \frac{n!}{2!(n-2)!}(n-2)! - \cdots + (-1)^n \frac{n!}{n!\,0!}0!. \end{aligned}$$

Simplifying this expression gives

$$D_n = n!\left[1 - \frac{1}{1!} + \frac{1}{2!} - \cdots + (-1)^n \frac{1}{n!}\right].$$

◁

It is now simple to find D_n for a given positive integer n. For instance, using Theorem 2, it follows that

$$D_3 = 3!\left[1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!}\right] = 6\left(1 - 1 + \frac{1}{2} - \frac{1}{6}\right) = 2,$$

as we have previously remarked.

The solution of the problem in Example 4 can now be given.

Solution: The probability that no one receives the correct hat is $D_n/n!$. By Theorem 2, this probability is

$$\frac{D_n}{n!} = 1 - \frac{1}{1!} + \frac{1}{2!} - \cdots + (-1)^n \frac{1}{n!}.$$

The values of this probability for $2 \le n \le 7$ are displayed in Table 2.

Using methods from calculus it can be shown that

$$e^{-1} = 1 - \frac{1}{1!} + \frac{1}{2!} - \cdots + (-1)^n \frac{1}{n!} + \cdots \sim 0.368.$$

Because this is an alternating series with terms tending to zero, it follows that as n grows without bound, the probability that no one receives the correct hat converges to $e^{-1} \sim 0.368$. In fact, this probability can be shown to be within $1/(n+1)!$ of e^{-1}. ◀

Exercises

1. Suppose that in a bushel of 100 apples there are 20 that have worms in them and 15 that have bruises. Only those apples with neither worms nor bruises can be sold. If there are 10 bruised apples that have worms in them, how many of the 100 apples can be sold?
2. Of 1000 applicants for a mountain-climbing trip in the Himalayas, 450 get altitude sickness, 622 are not in good enough shape, and 30 have allergies. An applicant qualifies if and only if this applicant does not get altitude sickness, is in good shape, and does not have allergies. If there are 111 applicants who get altitude sickness and are not in good enough shape, 14 who get altitude sickness and have allergies, 18 who are not in good enough shape and have allergies, and 9 who get altitude sickness, are

not in good enough shape, and have allergies, how many applicants qualify?

3. How many solutions does the equation $x_1 + x_2 + x_3 = 13$ have where x_1, x_2, and x_3 are nonnegative integers less than 6?
4. Find the number of solutions of the equation $x_1 + x_2 + x_3 + x_4 = 17$, where x_i, $i = 1, 2, 3, 4$, are nonnegative integers such that $x_1 \leq 3$, $x_2 \leq 4$, $x_3 \leq 5$, and $x_4 \leq 8$.
5. Find the number of primes less than 200 using the principle of inclusion–exclusion.
6. An integer is called **squarefree** if it is not divisible by the square of a positive integer greater than 1. Find the number of squarefree positive integers less than 100.
7. How many positive integers less than 10,000 are not the second or higher power of an integer?
8. How many onto functions are there from a set with seven elements to one with five elements?
9. How many ways are there to distribute six different toys to three different children such that each child gets at least one toy?
10. In how many ways can eight distinct balls be distributed into three distinct urns if each urn must contain at least one ball?
11. In how many ways can seven different jobs be assigned to four different employees so that each employee is assigned at least one job and the most difficult job is assigned to the best employee?
12. List all the derangements of $\{1, 2, 3, 4\}$.
13. How many derangements are there of a set with seven elements?
14. What is the probability that none of 10 people receives the correct hat if a hatcheck person hands their hats back randomly?
15. A machine that inserts letters into envelopes goes haywire and inserts letters randomly into envelopes. What is the probability that in a group of 100 letters
    a) no letter is put into the correct envelope?
    b) exactly one letter is put into the correct envelope?
    c) exactly 98 letters are put into the correct envelopes?
    d) exactly 99 letters are put into the correct envelopes?
    e) all letters are put into the correct envelopes?
16. A group of n students is assigned seats for each of two classes in the same classroom. How many ways can these seats be assigned if no student is assigned the same seat for both classes?

*17. How many ways can the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 be arranged so that no even digit is in its original position?

*18. Use a combinatorial argument to show that the sequence $\{D_n\}$, where D_n denotes the number of derangements of n objects, satisfies the recurrence relation

$$D_n = (n-1)(D_{n-1} + D_{n-2})$$

for $n \geq 2$. [*Hint:* Note that there are $n - 1$ choices for the first element k of a derangement. Consider separately the derangements that start with k that do and do not have 1 in the kth position.]

*19. Use Exercise 18 to show that

$$D_n = nD_{n-1} + (-1)^n$$

for $n \geq 1$.

20. Use Exercise 19 to find an explicit formula for D_n.
21. For which positive integers n is D_n, the number of derangements of n objects, even?
22. Suppose that p and q are distinct primes. Use the principle of inclusion–exclusion to find $\phi(pq)$, the number of positive integers not exceeding pq that are relatively prime to pq.

*23. Use the principle of inclusion–exclusion to derive a formula for $\phi(n)$ when the prime factorization of n is

$$n = p_1^{a_1} p_2^{a_2} \cdots p_m^{a_m}.$$

*24. Show that if n is a positive integer, then

$$n! = C(n, 0)D_n + C(n, 1)D_{n-1} + \cdots + C(n, n-1)D_1 + C(n, n)D_0,$$

where D_k is the number of derangements of k objects.

25. How many derangements of $\{1, 2, 3, 4, 5, 6\}$ begin with the integers 1, 2, and 3, in some order?
26. How many derangements of $\{1, 2, 3, 4, 5, 6\}$ end with the integers 1, 2, and 3, in some order?
27. Prove Theorem 1.

Key Terms and Results

TERMS

recurrence relation: a formula expressing terms of a sequence, except for some initial terms, as a function of one or more previous terms of the sequence

initial conditions for a recurrence relation: the values of the terms of a sequence satisfying the recurrence relation before this relation takes effect

linear homogeneous recurrence relation with constant coefficients: a recurrence relation that expresses the terms of a sequence, except initial terms, as a linear combination of previous terms

characteristic roots of a linear homogeneous recurrence relation with constant coefficients: the roots of the polynomial associated with a linear homogeneous recurrence relation with constant coefficients

linear nonhomogeneous recurrence relation with constant coefficients: a recurrence relation that expresses the terms of a sequence, except for initial terms, as a linear combination of previous terms plus a function that is not identically zero that depends only on the index

divide-and-conquer algorithm: an algorithm that solves a problem recursively by splitting it into a fixed number of smaller problems of the same type

generating function of a sequence: the formal series that has the nth term of the sequence as the coefficient of x^n

sieve of Eratosthenes: a procedure for finding the primes less than a specified positive integer

derangement: a permutation of objects such that no object is in its original place

RESULTS

the formula for the number of elements in the union of two finite sets:

$$|A \cup B| = |A| + |B| - |A \cap B|$$

the formula for the number of elements in the union of three finite sets:

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

the principle of inclusion–exclusion:

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = \sum_{1 \le i \le n} |A_i| - \sum_{1 \le i < j \le n} |A_i \cap A_j| + \sum_{1 \le i < j < k \le n} |A_i \cap A_j \cap A_k| - \cdots + (-1)^{n+1} |A_1 \cap A_2 \cap \cdots \cap A_n|$$

the number of onto functions from a set with m elements to a set with n elements:

$$n^m - C(n, 1)(n-1)^m + C(n, 2)(n-2)^m - \cdots + (-1)^{n-1} C(n, n-1) \cdot 1^m$$

the number of derangements of n objects:

$$D_n = n!\left[1 - \frac{1}{1!} + \frac{1}{2!} - \cdots + (-1)^n \frac{1}{n!}\right]$$

Review Questions

1. **a)** What is a recurrence relation?
   b) Find a recurrence relation for the amount of money that will be in an account after n years if \$1,000,000 is deposited in an account yielding 9% annually.
2. Explain how the Fibonacci numbers are used to solve Fibonacci's problem about rabbits.
3. **a)** Find a recurrence relation for the number of steps needed to solve the Tower of Hanoi puzzle.
   b) Show how this recurrence relation can be solved using iteration.
4. **a)** Explain how to find a recurrence relation for the number of bit strings of length n not containing two consecutive 1s.
   b) Describe another counting problem that has a solution satisfying the same recurrence relation.
5. Define a linear homogeneous recurrence relation of degree k.
6. **a)** Explain how to solve linear homogeneous recurrence relations of degree 2.
   b) Solve the recurrence relation $a_n = 13a_{n-1} - 22a_{n-2}$ for $n \ge 2$ if $a_0 = 3$ and $a_1 = 15$.
   c) Solve the recurrence relation $a_n = 14a_{n-1} - 49a_{n-2}$ for $n \ge 2$ if $a_0 = 3$ and $a_1 = 35$.
7. **a)** Explain how to find $f(b^k)$ where k is a positive integer if $f(n)$ satisfies the divide-and-conquer recurrence relation $f(n) = af(n/b) + g(n)$ whenever b divides the positive integer n.
   b) Find $f(256)$ if $f(n) = 3f(n/4) + 5n/4$ and $f(1) = 7$.
8. **a)** Derive a divide-and-conquer recurrence relation for the number of comparisons used to find a number in a list using a binary search.
   b) Give a big-O estimate for the number of comparisons used by a binary search from the divide-and-conquer recurrence relation you gave in (a) using Theorem 1 in Section 7.3.
9. **a)** Give a formula for the number of elements in the union of three sets.
   b) Explain why this formula is valid.
   c) Explain how to use the formula from (a) to find the number of integers not exceeding 1000 that are divisible by 6, 10, or 15.
   d) Explain how to use the formula from (a) to find the number of solutions in nonnegative integers to the equation $x_1 + x_2 + x_3 + x_4 = 22$ with $x_1 < 8$, $x_2 < 6$, and $x_3 < 5$.
10. **a)** Give a formula for the number of elements in the union of four sets and explain why it is valid.

b) Suppose the sets A_1, A_2, A_3, and A_4 each contain 25 elements, the intersection of any two of these sets contains 5 elements, the intersection of any three of these sets contains 2 elements, and 1 element is in all four of the sets. How many elements are in the union of the four sets?

11. a) State the principle of inclusion–exclusion.
b) Outline a proof of this principle.

12. Explain how the principle of inclusion–exclusion can be used to count the number of onto functions from a set with m elements to a set with n elements.

13. a) How can you count the number of ways to assign m jobs to n employees so that each employee is assigned at least one job?
b) How many ways are there to assign seven jobs to three employees so that each employee is assigned at least one job?

14. Explain how the inclusion–exclusion principle can be used to count the number of primes not exceeding the positive integer n.

15. a) Define a derangement.
b) Why is counting the number of ways a hatcheck person can return hats to n people, so that no one receives the correct hat, the same as counting the number of derangements of n objects?
c) Explain how to count the number of derangements of n objects.

Supplementary Exercises

1. A group of 10 people begin a chain letter, with each person sending the letter to four other people. Each of these people sends the letter to four additional people.
a) Find a recurrence relation for the number of letters sent at the nth stage of this chain letter, if no person ever receives more than one letter.
b) What are the initial conditions for the recurrence relation in part (a)?
c) How many letters are sent at the nth stage of the chain letter?

2. A nuclear reactor has created 18 grams of a particular radioactive isotope. Every hour 1% of a radioactive isotope decays.
a) Set up a recurrence relation for the amount of this isotope left after n hours.
b) What are the initial conditions for the recurrence relation in part (a)?
c) Solve this recurrence relation.

3. Every hour the U.S. government prints 10,000 more \$1 bills, 4000 more \$5 bills, 3000 more \$10 bills, 2500 more \$20 bills, 1000 more \$50 bills, and the same number of \$100 bills as it did the previous hour. In the initial hour 1000 of each bill were produced.
a) Set up a recurrence relation for the amount of money produced in the nth hour.
b) What are the initial conditions for the recurrence relation in part (a)?
c) Solve the recurrence relation for the amount of money produced in the nth hour.
d) Set up a recurrence relation for the total amount of money produced in the first n hours.
e) Solve the recurrence relation for the total amount of money produced in the first n hours.

4. Suppose that every hour there are two new bacteria in a colony for each bacterium that was present the previous hour, and that all bacteria 2 hours old die. The colony starts with 100 new bacteria.
a) Set up a recurrence relation for the number of bacteria present after n hours.
b) What is the solution of this recurrence relation?
c) When will the colony contain more than 1 million bacteria?

5. Messages are sent over a communications channel using two different signals. One signal requires 2 microseconds for transmittal, and the other signal requires 3 microseconds for transmittal. Each signal of a message is followed immediately by the next signal.
a) Find a recurrence relation for the number of different signals that can be sent in n microseconds.
b) What are the initial conditions of the recurrence relation in part (a)?
c) How many different messages can be sent in 12 microseconds?

6. A small post office has only 4-cent stamps, 6-cent stamps, and 10-cent stamps. Find a recurrence relation for the number of ways to form postage of n cents with these stamps if the order that the stamps are used matters. What are the initial conditions for this recurrence relation?

7. How many ways are there to form these postages using the rules described in Exercise 6?
a) 12 cents b) 14 cents
c) 18 cents d) 22 cents

8. Find the solutions of the simultaneous system of congruences

$$a_n = a_{n-1} + b_{n-1}$$
$$b_n = a_{n-1} - b_{n-1}$$

with $a_0 = 1$ and $b_0 = 2$.

9. Solve the recurrence relation $a_n = a_{n-1}^2/a_{n-2}$ if $a_0 = 1$ and $a_1 = 2$. [*Hint:* Take logarithms of both sides to obtain a recurrence relation for the sequence $\log a_n$, $n = 0, 1, 2, \ldots$.]

*10. Solve the recurrence relation $a_n = a_{n-1}^3 a_{n-2}^2$ if $a_0 = 2$ and $a_1 = 2$. (See the hint for Exercise 9.)

11. Find the solution of the recurrence relation $a_n = 3a_{n-1} - 3a_{n-2} + a_{n-3} + 1$ if $a_0 = 2$, $a_1 = 4$, and $a_2 = 8$.

12. Find the solution of the recurrence relation $a_n = 3a_{n-1} - 3a_{n-2} + a_{n-3}$ if $a_0 = 2$, $a_1 = 2$, and $a_2 = 4$.

*13. Suppose that in Example 4 of Section 7.1 a pair of rabbits leaves the island after reproducing twice. Find a recurrence relation for the number of rabbits on the island in the middle of the nth month.

14. Find the solution to the recurrence relation $f(n) = 3f(n/5) + 2n^4$, when n is divisible by 5, for $n = 5^k$, where k is a positive integer and $f(1) = 1$.

15. Estimate the size of f in Exercise 14 if f is an increasing function.

16. Find a recurrence relation that describes the number of comparisons used by the following algorithm: Find the largest and second largest elements of a sequence of n numbers recursively by splitting the sequence into two subsequences with an equal number of terms, or where there is one more term in one subsequence than in the other, at each stage. Stop when subsequences with two terms are reached.

17. Estimate the number of comparisons used by the algorithm described in Exercise 16.

Let $\{a_n\}$ be a sequence of real numbers. The **forward differences** of this sequence are defined recursively as follows: The **first forward difference** is $\Delta a_n = a_{n+1} - a_n$; the **$(k+1)$st forward difference** $\Delta^{k+1} a_n$ is obtained from $\Delta^k a_n$ by $\Delta^{k+1} a_n = \Delta^k a_{n+1} - \Delta^k a_n$.

18. Find Δa_n, where

a) $a_n = 3$. **b)** $a_n = 4n + 7$. **c)** $a_n = n^2 + n + 1$.

19. Let $a_n = 3n^3 + n + 2$. Find $\Delta^k a_n$, where k equals

a) 2. **b)** 3. **c)** 4.

*20. Suppose that $a_n = P(n)$, where P is a polynomial of degree d. Prove that $\Delta^{d+1} a_n = 0$ for all nonnegative integers n.

21. Let $\{a_n\}$ and $\{b_n\}$ be sequences of real numbers. Show that

$$\Delta(a_n b_n) = a_{n+1}(\Delta b_n) + b_n(\Delta a_n).$$

22. Show that if $F(x)$ and $G(x)$ are the generating functions for the sequences $\{a_k\}$ and $\{b_k\}$, respectively, and c and d are real numbers, then $(cF + dG)(x)$ is the generating function for $\{ca_k + db_k\}$.

23. (*Requires calculus*) This exercise shows how generating functions can be used to solve the recurrence relation $(n+1)a_{n+1} = a_n + (1/n!)$ for $n \geq 0$ with initial condition $a_0 = 1$.

a) Let $G(x)$ be the generating function for $\{a_n\}$. Show that $G'(x) = G(x) + e^x$ and $G(0) = 1$.

b) Show from part (a) that $(e^{-x}G(x))' = 1$, and conclude that $G(x) = xe^x + e^x$.

c) Use part (b) to find a closed form for a_n.

24. Suppose that 14 students receive an A on the first exam in a discrete mathematics class, and 18 receive an A on the second exam. If 22 students received an A on either the first exam or the second exam, how many students received an A on both exams?

25. There are 323 farms in Monmouth County that have at least one of horses, cows, and sheep. If 224 have horses, 85 have cows, 57 have sheep, and 18 farms have all three types of animals, how many farms have exactly two of these three types of animals?

26. Queries to a database of student records at a college produced the following data: There are 2175 students at the college, 1675 of these are not freshmen, 1074 students have taken a course in calculus, 444 students have taken a course in discrete mathematics, 607 students are not freshmen and have taken calculus, 350 students have taken calculus and discrete mathematics, 201 students are not freshmen and have taken discrete mathematics, and 143 students are not freshmen and have taken both calculus and discrete mathematics. Can all the responses to the queries be correct?

27. Students in the school of mathematics at a university major in one or more of the following four areas: applied mathematics (AM), pure mathematics (PM), operations research (OR), and computer science (CS). How many students are in this school if (including joint majors) there are 23 students majoring in AM; 17 students majoring in PM; 44 in OR; 63 in CS; 5 in AM and PM; 8 in AM and CS; 4 in AM and OR; 6 in PM and CS; 5 in PM and OR; 14 in OR and CS; 2 in PM, OR, and CS; 2 in AM, OR, and CS; 1 in PM, AM, and OR; 1 in PM, AM, and CS; and 1 in all four fields.

28. How many terms are needed when the inclusion–exclusion principle is used to express the number of elements in the union of seven sets if no more than five of these sets have a common element?

29. How many solutions in positive integers are there to the equation $x_1 + x_2 + x_3 = 20$ with $2 < x_1 < 6$, $6 < x_2 < 10$, and $0 < x_3 < 5$?

30. How many positive integers less than 1,000,000 are

a) divisible by 2, 3, or 5?

b) not divisible by 7, 11, or 13?

c) divisible by 3 but not by 7?

31. How many positive integers less than 200 are

a) second or higher powers of integers?

b) either second or higher powers of integers or primes?

c) not divisible by the square of an integer greater than 1?

d) not divisible by the cube of an integer greater than 1?

e) not divisible by three or more primes?

*32. How many ways are there to assign six different jobs to three different employees if the hardest job is assigned

to the most experienced employee and the easiest job is assigned to the least experienced employee?

33. What is the probability that exactly one person is given back the correct hat by a hatcheck person who gives n people their hats back at random?

34. How many bit strings of length six do not contain four consecutive 1s?

35. What is the probability that a bit string of length six contains at least four 1s?

Computer Projects

Write programs with these input and output.

1. Given a positive integer n, list all the moves required in the Tower of Hanoi puzzle to move n disks from one peg to another according to the rules of the puzzle.
2. Given a positive integer n and an integer k with $1 \leq k \leq n$, list all the moves used by the Frame–Stewart algorithm (described in the preamble to Exercise 54 of Section 7.1) to move n disks from one peg to another using four pegs according to the rules of the puzzle.
3. Given a positive integer n, list all the bit sequences of length n that do not have a pair of consecutive 0s.
4. Given a positive integer n, write out all ways to parenthesize the product of $n + 1$ variables.
5. Given a recurrence relation $a_n = c_1 a_{n-1} + c_2 a_{n-2}$, where c_1 and c_2 are real numbers, initial conditions $a_0 = C_0$ and $a_1 = C_1$, and a positive integer k, find a_k using iteration.
6. Given a recurrence relation $a_n = c_1 a_{n-1} + c_2 a_{n-2}$ and initial conditions $a_0 = C_0$ and $a_1 = C_1$, determine the unique solution.
7. Given a recurrence relation of the form $f(n) = af(n/b) + c$, where a is a real number, b is a positive integer, and c is a real number, and a positive integer k, find $f(b^k)$ using iteration.
8. Given the number of elements in the intersection of three sets, the number of elements in each pairwise intersection of these sets, and the number of elements in each set, find the number of elements in their union.
9. Given a positive integer n, produce the formula for the number of elements in the union of n sets.
10. Given positive integers m and n, find the number of onto functions from a set with m elements to a set with n elements.
11. Given a positive integer n, list all the derangements of the set $\{1, 2, 3, \ldots, n\}$.

Computations and Explorations

Use a computational program or programs you have written to do these exercises.

1. Find the exact value of f_{100}, f_{500}, and f_{1000}, where f_n is the nth Fibonacci number.
2. Find the smallest Fibonacci number greater than 1,000,000, greater than 1,000,000,000, and greater than 1,000,000,000,000.
3. Find as many prime Fibonacci numbers as you can. It is unknown whether there are infinitely many of these.
4. Write out all the moves required to solve the Tower of Hanoi puzzle with 10 disks.
5. Write out all the moves required to use the Frame–Stewart algorithm to move 20 disks from one peg to another peg using four pegs according to the rules of the Reve's puzzle.
6. Verify the Frame conjecture for solving the Reve's puzzle for n disks for as many integers n as possible by showing that the puzzle cannot be solved using fewer moves than are made by the Frame–Stewart algorithm with the optimal choice of k.
7. Compute the number of operations required to multiply two integers with n bits for various integers n including 16, 64, 256, and 1024 using the fast multiplication described in Section 7.3 and the standard algorithm for multiplying integers (Algorithm 3 in Section 3.6).
8. Compute the number of operations required to multiply two $n \times n$ matrices for various integers n including 4, 16, 64, and 128 using the fast matrix multiplication described in Section 7.3 and the standard algorithm for multiplying matrices (Algorithm 1 in Section 3.8).
9. Use the sieve of Eratosthenes to find all the primes not exceeding 1000.
10. Find the number of primes not exceeding 10,000 using the method described in Section 7.6 to find the number of primes not exceeding 100.
11. List all the derangements of $\{1, 2, 3, 4, 5, 6, 7, 8\}$.
12. Compute the probability that a permutation of n objects is a derangement for all positive integers not exceeding 20 and determine how quickly these probabilities approach the number e.

Writing Projects

Respond with essays using outside sources.

1. Find the original source where Fibonacci presented his puzzle about modeling rabbit populations. Discuss this problem and other problems posed by Fibonacci and give some information about Fibonacci himself.
2. Explain how the Fibonacci numbers arise in a variety of applications, such as in phyllotaxis, the study of arrangement of leaves in plants, in the study of reflections by mirrors, and so on.
3. Describe different variations of the Tower of Hanoi puzzle, including those with more than three pegs (including the Reve's puzzle discussed in the text and exercises), those where disk moves are restricted, and those where disks may have the same size. Include what is known about the number of moves required to solve each variation.
4. Discuss as many different problems as possible where the Catalan numbers arise.
5. Describe the solution of Ulam's problem (see Exercise 28 in Section 7.3) involving searching with one lie found by Andrzej Pelc.
6. Discuss variations of Ulam's problem (see Exercise 28 in Section 7.3) involving searching with more than one lie and what is known about this problem.
7. Define the convex hull of a set of points in the plane and describe divide-and-conquer algorithms for finding the convex hull of a set of points in the plane.
8. Look up the definition of the *lucky numbers*. Explain how they are found using a sieve technique similar to the sieve of Eratosthenes. Find all the lucky numbers less than 1000.
9. Describe how sieve methods are used in number theory. What kind of results have been established using such methods?
10. Look up the rules of the old French card game of *rencontres*. Describe these rules and describe the work of Pierre Raymond de Montmort on *le problème de rencontres*.
11. Describe how exponential generating functions can be used to solve a variety of counting problems.
12. Describe the Polyá theory of counting and the kind of counting problems that can be solved using this theory.
13. The *problème des ménages* (the problem of the households) asks for the number of ways to arrange n couples around a table so that the sexes alternate and no husband and wife are seated together. Explain the method used by E. Lucas to solve this problem.
14. Explain how *rookpolynomials* can be used to solve counting problems.

CHAPTER

10 Trees

10.1 Introduction to Trees

10.2 Applications of Trees

10.3 Tree Traversal

10.4 Spanning Trees

10.5 Minimum Spanning Trees

A connected graph that contains no simple circuits is called a tree. Trees were used as long ago as 1857, when the English mathematician Arthur Cayley used them to count certain types of chemical compounds. Since that time, trees have been employed to solve problems in a wide variety of disciplines, as the examples in this chapter will show.

Trees are particularly useful in computer science, where they are employed in a wide range of algorithms. For instance, trees are used to construct efficient algorithms for locating items in a list. They can be used in algorithms, such as Huffman coding, that construct efficient codes saving costs in data transmission and storage. Trees can be used to study games such as checkers and chess and can help determine winning strategies for playing these games. Trees can be used to model procedures carried out using a sequence of decisions. Constructing these models can help determine the computational complexity of algorithms based on a sequence of decisions, such as sorting algorithms.

Procedures for building trees containing every vertex of a graph, including depth-first search and breadth-first search, can be used to systematically explore the vertices of a graph. Exploring the vertices of a graph via depth-first search, also known as backtracking, allows for the systematic search for solutions to a wide variety of problems, such as determining how eight queens can be placed on a chessboard so that no queen can attack another.

We can assign weights to the edges of a tree to model many problems. For example, using weighted trees we can develop algorithms to construct networks containing the least expensive set of telephone lines linking different network nodes.

10.1 Introduction to Trees

Limits

In Chapter 9 we showed how graphs can be used to model and solve many problems. In this chapter we will focus on a particular type of graph called a **tree,** so named because such graphs resemble trees. For example, *family trees* are graphs that represent genealogical charts. Family trees use vertices to represent the members of a family and edges to represent parent–child relationships. The family tree of the male members of the Bernoulli family of Swiss mathematicians is shown in Figure 1. The undirected graph representing a family tree (restricted to people of just one gender and with no inbreeding) is an example of a tree.

DEFINITION 1 A *tree* is a connected undirected graph with no simple circuits.

Because a tree cannot have a simple circuit, a tree cannot contain multiple edges or loops. Therefore any tree must be a simple graph.

EXAMPLE 1 Which of the graphs shown in Figure 2 are trees?

Solution: G_1 and G_2 are trees, because both are connected graphs with no simple circuits. G_3 is not a tree because e, b, a, d, e is a simple circuit in this graph. Finally, G_4 is not a tree because it is not connected. ◀

FIGURE 1 The Bernoulli Family of Mathematicians.

Any connected graph that contains no simple circuits is a tree. What about graphs containing no simple circuits that are not necessarily connected? These graphs are called **forests** and have the property that each of their connected components is a tree. Figure 3 displays a forest.

Trees are often defined as undirected graphs with the property that there is a unique simple path between every pair of vertices. Theorem 1 shows that this alternative definition is equivalent to our definition.

THEOREM 1 An undirected graph is a tree if and only if there is a unique simple path between any two of its vertices.

Proof: First assume that T is a tree. Then T is a connected graph with no simple circuits. Let x and y be two vertices of T. Because T is connected, by Theorem 1 of Section 9.4 there is a simple path between x and y. Moreover, this path must be unique, for if there were a second such path, the path formed by combining the first path from x to y followed by the path from y to x obtained by reversing the order of the second path from x to y would form a circuit. This

FIGURE 2 Examples of Trees and Graphs That Are Not Trees.

FIGURE 3 Example of a Forest.

implies, using Exercise 49 of Section 9.4, that there is a simple circuit in T. Hence, there is a unique simple path between any two vertices of a tree.

Now assume that there is a unique simple path between any two vertices of a graph T. Then T is connected, because there is a path between any two of its vertices. Furthermore, T can have no simple circuits. To see that this is true, suppose T had a simple circuit that contained the vertices x and y. Then there would be two simple paths between x and y, because the simple circuit is made up of a simple path from x to y and a second simple path from y to x. Hence, a graph with a unique simple path between any two vertices is a tree. ◁

In many applications of trees, a particular vertex of a tree is designated as the **root.** Once we specify a root, we can assign a direction to each edge as follows. Because there is a unique path from the root to each vertex of the graph (by Theorem 1), we direct each edge away from the root. Thus, a tree together with its root produces a directed graph called a **rooted tree.**

DEFINITION 2 A *rooted tree* is a tree in which one vertex has been designated as the root and every edge is directed away from the root.

Rooted trees can also be defined recursively. Refer to Section 4.3 to see how this can be done. We can change an unrooted tree into a rooted tree by choosing any vertex as the root. Note that different choices of the root produce different rooted trees. For instance, Figure 4 displays the rooted trees formed by designating a to be the root and c to be the root, respectively, in the tree T. We usually draw a rooted tree with its root at the top of the graph. The arrows indicating the

FIGURE 4 A Tree and Rooted Trees Formed by Designating Two Roots.

FIGURE 5 A Rooted Tree T.

FIGURE 6 The Subtree Rooted at g.

directions of the edges in a rooted tree can be omitted, because the choice of root determines the directions of the edges.

The terminology for trees has botanical and genealogical origins. Suppose that T is a rooted tree. If v is a vertex in T other than the root, the **parent** of v is the unique vertex u such that there is a directed edge from u to v (the reader should show that such a vertex is unique). When u is the parent of v, v is called a **child** of u. Vertices with the same parent are called **siblings.** The **ancestors** of a vertex other than the root are the vertices in the path from the root to this vertex, excluding the vertex itself and including the root (that is, its parent, its parent's parent, and so on, until the root is reached). The **descendants** of a vertex v are those vertices that have v as an ancestor. A vertex of a tree is called a **leaf** if it has no children. Vertices that have children are called **internal vertices.** The root is an internal vertex unless it is the only vertex in the graph, in which case it is a leaf.

If a is a vertex in a tree, the **subtree** with a as its root is the subgraph of the tree consisting of a and its descendants and all edges incident to these descendants.

EXAMPLE 2 In the rooted tree T (with root a) shown in Figure 5, find the parent of c, the children of g, the siblings of h, all ancestors of e, all descendants of b, all internal vertices, and all leaves. What is the subtree rooted at g?

Extra Examples

Solution: The parent of c is b. The children of g are h, i, and j. The siblings of h are i and j. The ancestors of e are c, b, and a. The descendants of b are c, d, and e. The internal vertices are a, b, c, g, h, and j. The leaves are d, e, f, i, k, l, and m. The subtree rooted at g is shown in Figure 6. ◀

Rooted trees with the property that all of their internal vertices have the same number of children are used in many different applications. Later in this chapter we will use such trees to study problems involving searching, sorting, and coding.

DEFINITION 3 A rooted tree is called an *m-ary tree* if every internal vertex has no more than m children. The tree is called a *full m-ary tree* if every internal vertex has exactly m children. An m-ary tree with $m = 2$ is called a *binary tree.*

Links

FIGURE 7 Four Rooted Trees.

EXAMPLE 3 Are the rooted trees in Figure 7 full m-ary trees for some positive integer m?

Solution: T_1 is a full binary tree because each of its internal vertices has two children. T_2 is a full 3-ary tree because each of its internal vertices has three children. In T_3 each internal vertex has five children, so T_3 is a full 5-ary tree. T_4 is not a full m-ary tree for any m because some of its internal vertices have two children and others have three children. ◀

An **ordered rooted tree** is a rooted tree where the children of each internal vertex are ordered. Ordered rooted trees are drawn so that the children of each internal vertex are shown in order from left to right. Note that a representation of a rooted tree in the conventional way determines an ordering for its edges. We will use such orderings of edges in drawings without explicitly mentioning that we are considering a rooted tree to be ordered.

In an ordered binary tree (usually called just a **binary tree**), if an internal vertex has two children, the first child is called the **left child** and the second child is called the **right child.** The tree rooted at the left child of a vertex is called the **left subtree** of this vertex, and the tree rooted at the right child of a vertex is called the **right subtree** of the vertex. The reader should note that for some applications every vertex of a binary tree, other than the root, is designated as a right or a left child of its parent. This is done even when some vertices have only one child. We will make such designations whenever it is necessary, but not otherwise.

Ordered rooted trees can be defined recursively. Binary trees, a type of ordered rooted trees, were defined this way in Section 4.3.

EXAMPLE 4 What are the left and right children of d in the binary tree T shown in Figure 8(a) (where the order is that implied by the drawing)? What are the left and right subtrees of c?

Solution: The left child of d is f and the right child is g. We show the left and right subtrees of c in Figures 8(b) and 8(c), respectively. ◀

FIGURE 8 A Binary Tree T and Left and Right Subtrees of the Vertex c.

Butane Isobutane

FIGURE 9 The Two Isomers of Butane.

Just as in the case of graphs, there is no standard terminology used to describe trees, rooted trees, ordered rooted trees, and binary trees. This nonstandard terminology occurs because trees are used extensively throughout computer science, which is a relatively young field. The reader should carefully check meanings given to terms dealing with trees whenever they occur.

Trees as Models

Trees are used as models in such diverse areas as computer science, chemistry, geology, botany, and psychology. We will describe a variety of such models based on trees.

EXAMPLE 5 **Saturated Hydrocarbons and Trees** Graphs can be used to represent molecules, where atoms are represented by vertices and bonds between them by edges. The English mathematician Arthur Cayley discovered trees in 1857 when he was trying to enumerate the isomers of compounds of the form C_nH_{2n+2}, which are called *saturated hydrocarbons.*

In graph models of saturated hydrocarbons, each carbon atom is represented by a vertex of degree 4, and each hydrogen atom is represented by a vertex of degree 1. There are $3n + 2$ vertices in a graph representing a compound of the form C_nH_{2n+2}. The number of edges in such a graph is half the sum of the degrees of the vertices. Hence, there are $(4n + 2n + 2)/2 = 3n + 1$ edges in this graph. Because the graph is connected and the number of edges is one less than the number of vertices, it must be a tree (see Exercise 15 at the end of this section).

The nonisomorphic trees with n vertices of degree 4 and $2n + 2$ of degree 1 represent the different isomers of C_nH_{2n+2}. For instance, when $n = 4$, there are exactly two nonisomorphic trees of this type (the reader should verify this). Hence, there are exactly two different isomers of C_4H_{10}. Their structures are displayed in Figure 9. These two isomers are called butane and isobutane. ◀

Links

ARTHUR CAYLEY (1821–1895) Arthur Cayley, the son of a merchant, displayed his mathematical talents at an early age with amazing skill in numerical calculations. Cayley entered Trinity College, Cambridge, when he was 17. While in college he developed a passion for reading novels. Cayley excelled at Cambridge and was elected to a 3-year appointment as Fellow of Trinity and assistant tutor. During this time Cayley began his study of n-dimensional geometry and made a variety of contributions to geometry and to analysis. He also developed an interest in mountaineering, which he enjoyed during vacations in Switzerland. Because no position as a mathematician was available to him, Cayley left Cambridge, entering the legal profession and gaining admittance to the bar in 1849. Although Cayley limited his legal work to be able to continue his mathematics research, he developed a reputation as a legal specialist. During his legal career he was able to write more than 300 mathematical papers. In 1863 Cambridge University established a new post in mathematics and offered it to Cayley. He took this job, even though it paid less money than he made as a lawyer.

FIGURE 10 An Organizational Tree for a Computer Company.

EXAMPLE 6 **Representing Organizations** The structure of a large organization can be modeled using a rooted tree. Each vertex in this tree represents a position in the organization. An edge from one vertex to another indicates that the person represented by the initial vertex is the (direct) boss of the person represented by the terminal vertex. The graph shown in Figure 10 displays such a tree. In the organization represented by this tree, the Director of Hardware Development works directly for the Vice President of R&D. ◀

EXAMPLE 7 **Computer File Systems** Files in computer memory can be organized into directories. A directory can contain both files and subdirectories. The root directory contains the entire file system. Thus, a file system may be represented by a rooted tree, where the root represents the root directory, internal vertices represent subdirectories, and leaves represent ordinary files or empty directories. One such file system is shown in Figure 11. In this system, the file khr is in the directory rje. (Note that links to files where the same file may have more than one pathname can lead to circuits in computer file systems.) ◀

EXAMPLE 8 **Tree-Connected Parallel Processors** In Example 16 of Section 9.2 we described several interconnection networks for parallel processing. A **tree-connected network** is another important

FIGURE 11 A Computer File System.

way to interconnect processors. The graph representing such a network is a complete binary tree. Such a network interconnects $n = 2^k - 1$ processors, where k is a positive integer. A processor represented by the vertex v that is not a root or a leaf has three two-way connections—one to the processor represented by the parent of v and two to the processors represented by the two children of v. The processor represented by the root has two two-way connections to the processors represented by its two children. A processor represented by a leaf v has a single two-way connection to the parent of v. We display a tree-connected network with seven processors in Figure 12.

FIGURE 12 A Tree-Connected Network of Seven Processors.

We will illustrate how a tree-connected network can be used for parallel computation. In particular, we will show how the processors in Figure 12 can be used to add eight numbers, using three steps. In the first step, we add x_1 and x_2 using P_4, x_3 and x_4 using P_5, x_5 and x_6 using P_6, and x_7 and x_8 using P_7. In the second step, we add $x_1 + x_2$ and $x_3 + x_4$ using P_2 and $x_5 + x_6$ and $x_7 + x_8$ using P_3. Finally, in the third step, we add $x_1 + x_2 + x_3 + x_4$ and $x_5 + x_6 + x_7 + x_8$ using P_1. The three steps used to add eight numbers compares favorably to the seven steps required to add eight numbers serially, where the steps are the addition of one number to the sum of the previous numbers in the list. ◀

Properties of Trees

We will often need results relating the numbers of edges and vertices of various types in trees.

THEOREM 2 A tree with n vertices has $n - 1$ edges.

Links

Proof: We will use mathematical induction to prove this theorem. Note that for all the trees here we can choose a root and consider the tree rooted.

BASIS STEP: When $n = 1$, a tree with $n = 1$ vertex has no edges. It follows that the theorem is true for $n = 1$.

INDUCTIVE STEP: The induction hypothesis states that every tree with k vertices has $k - 1$ edges, where k is a positive integer. Suppose that a tree T has $k + 1$ vertices and that v is a leaf of T (which must exist because the tree is finite), and let w be the parent of v. Removing from T the vertex v and the edge connecting w to v produces a tree T' with k vertices, because the resulting graph is still connected and has no simple circuits. By the induction hypothesis, T' has $k - 1$ edges. It follows that T has k edges because it has one more edge than T', the edge connecting v and w. This completes the induction step. ◁

The number of vertices in a full m-ary tree with a specified number of internal vertices is determined, as Theorem 3 shows. As in Theorem 2, we will use n to denote the number of vertices in a tree.

THEOREM 3 A full m-ary tree with i internal vertices contains $n = mi + 1$ vertices.

Proof: Every vertex, except the root, is the child of an internal vertex. Because each of the i internal vertices has m children, there are mi vertices in the tree other than the root. Therefore, the tree contains $n = mi + 1$ vertices. ◁

Suppose that T is a full m-ary tree. Let i be the number of internal vertices and l the number of leaves in this tree. Once one of n, i, and l is known, the other two quantities are determined. Theorem 4 explains how to find the other two quantities from the one that is known.

THEOREM 4 A full m-ary tree with

(*i*) n vertices has $i = (n-1)/m$ internal vertices and $l = [(m-1)n+1]/m$ leaves,
(*ii*) i internal vertices has $n = mi + 1$ vertices and $l = (m-1)i + 1$ leaves,
(*iii*) l leaves has $n = (ml-1)/(m-1)$ vertices and $i = (l-1)/(m-1)$ internal vertices.

Proof: Let n represent the number of vertices, i the number of internal vertices, and l the number of leaves. The three parts of the theorem can all be proved using the equality given in Theorem 3, that is, $n = mi + 1$, together with the equality $n = l + i$, which is true because each vertex is either a leaf or an internal vertex. We will prove part (*i*) here. The proofs of parts (*ii*) and (*iii*) are left as exercises for the reader.

Solving for i in $n = mi + 1$ gives $i = (n-1)/m$. Then inserting this expression for i into the equation $n = l + i$ shows that $l = n - i = n - (n-1)/m = [(m-1)n+1]/m$. ◁

Example 9 illustrates how Theorem 4 can be used.

EXAMPLE 9 Suppose that someone starts a chain letter. Each person who receives the letter is asked to send it on to four other people. Some people do this, but others do not send any letters. How many people have seen the letter, including the first person, if no one receives more than one letter and if the chain letter ends after there have been 100 people who read it but did not send it out? How many people sent out the letter?

Solution: The chain letter can be represented using a 4-ary tree. The internal vertices correspond to people who sent out the letter, and the leaves correspond to people who did not send it out. Because 100 people did not send out the letter, the number of leaves in this rooted tree is $l = 100$. Hence, part (*iii*) of Theorem 4 shows that the number of people who have seen the letter is $n = (4 \cdot 100 - 1)/(4-1) = 133$. Also, the number of internal vertices is $133 - 100 = 33$, so 33 people sent out the letter. ◀

FIGURE 13 A Rooted Tree.

It is often desirable to use rooted trees that are "balanced" so that the subtrees at each vertex contain paths of approximately the same length. Some definitions will make this concept clear. The **level** of a vertex v in a rooted tree is the length of the unique path from the root to this vertex. The level of the root is defined to be zero. The **height** of a rooted tree is the maximum of the levels of vertices. In other words, the height of a rooted tree is the length of the longest path from the root to any vertex.

EXAMPLE 10 Find the level of each vertex in the rooted tree shown in Figure 13. What is the height of this tree?

Solution: The root a is at level 0. Vertices b, j, and k are at level 1. Vertices c, e, f, and l are at level 2. Vertices d, g, i, m, and n are at level 3. Finally, vertex h is at level 4. Because the largest level of any vertex is 4, this tree has height 4. ◀

A rooted m-ary tree of height h is **balanced** if all leaves are at levels h or $h-1$.

FIGURE 14 Some Rooted Trees.

EXAMPLE 11 Which of the rooted trees shown in Figure 14 are balanced?

Solution: T_1 is balanced, because all its leaves are at levels 3 and 4. However, T_2 is not balanced, because it has leaves at levels 2, 3, and 4. Finally, T_3 is balanced, because all its leaves are at level 3. ◀

The results in Theorem 5 relate the height and the number of leaves in m-ary trees.

THEOREM 5 There are at most m^h leaves in an m-ary tree of height h.

Proof: The proof uses mathematical induction on the height. First, consider m-ary trees of height 1. These trees consist of a root with no more than m children, each of which is a leaf. Hence, there are no more than $m^1 = m$ leaves in an m-ary tree of height 1. This is the basis step of the inductive argument.

Now assume that the result is true for all m-ary trees of height less than h; this is the inductive hypothesis. Let T be an m-ary tree of height h. The leaves of T are the leaves of the subtrees of T obtained by deleting the edges from the root to each of the vertices at level 1, as shown in Figure 15.

Each of these subtrees has height less than or equal to $h - 1$. So by the inductive hypothesis, each of these rooted trees has at most m^{h-1} leaves. Because there are at most m such subtrees, each with a maximum of m^{h-1} leaves, there are at most $m \cdot m^{h-1} = m^h$ leaves in the rooted tree. This finishes the inductive argument. ◁

FIGURE 15 The Inductive Step of the Proof.

COROLLARY 1 If an m-ary tree of height h has l leaves, then $h \geq \lceil \log_m l \rceil$. If the m-ary tree is full and balanced, then $h = \lceil \log_m l \rceil$. (We are using the ceiling function here. Recall that $\lceil x \rceil$ is the smallest integer greater than or equal to x.)

Proof: We know that $l \leq m^h$ from Theorem 5. Taking logarithms to the base m shows that $\log_m l \leq h$. Because h is an integer, we have $h \geq \lceil \log_m l \rceil$. Now suppose that the tree is balanced. Then each leaf is at level h or $h - 1$, and because the height is h, there is at least one leaf at level h. It follows that there must be more than m^{h-1} leaves (see Exercise 30 at the end of this section). Because $l \leq m^h$, we have $m^{h-1} < l \leq m^h$. Taking logarithms to the base m in this inequality gives $h - 1 < \log_m l \leq h$. Hence, $h = \lceil \log_m l \rceil$. ◁

Exercises

1. Which of these graphs are trees?

2. Which of these graphs are trees?

3. Answer these questions about the rooted tree illustrated.

a) Which vertex is the root?
b) Which vertices are internal?
c) Which vertices are leaves?
d) Which vertices are children of j?
e) Which vertex is the parent of h?
f) Which vertices are siblings of o?
g) Which vertices are ancestors of m?
h) Which vertices are descendants of b?

4. Answer the same questions as listed in Exercise 3 for the rooted tree illustrated.

5. Is the rooted tree in Exercise 3 a full m-ary tree for some positive integer m?
6. Is the rooted tree in Exercise 4 a full m-ary tree for some positive integer m?
7. What is the level of each vertex of the rooted tree in Exercise 3?
8. What is the level of each vertex of the rooted tree in Exercise 4?
9. Draw the subtree of the tree in Exercise 3 that is rooted at
   a) a. **b)** c. **c)** e.
10. Draw the subtree of the tree in Exercise 4 that is rooted at
   a) a. **b)** c. **c)** e.
11. **a)** How many nonisomorphic unrooted trees are there with three vertices?
   b) How many nonisomorphic rooted trees are there with three vertices (using isomorphism for directed graphs)?

*12. **a)** How many nonisomorphic unrooted trees are there with four vertices?
   b) How many nonisomorphic rooted trees are there with four vertices (using isomorphism for directed graphs)?

*13. **a)** How many nonisomorphic unrooted trees are there with five vertices?
   b) How many nonisomorphic rooted trees are there with five vertices (using isomorphism for directed graphs)?

*14. Show that a simple graph is a tree if and only if it is connected, but the deletion of any of its edges produces a graph that is not connected.

*15. Let G be a simple graph with n vertices. Show that G is a tree if and only if G is connected and has $n - 1$ edges. [*Hint:* For the "if" part, use Exercise 14 and Theorem 2.]

16. Which complete bipartite graphs $K_{m,n}$, where m and n are positive integers, are trees?
17. How many edges does a tree with 10,000 vertices have?
18. How many vertices does a full 5-ary tree with 100 internal vertices have?
19. How many edges does a full binary tree with 1000 internal vertices have?
20. How many leaves does a full 3-ary tree with 100 vertices have?
21. Suppose 1000 people enter a chess tournament. Use a rooted tree model of the tournament to determine how many games must be played to determine a champion, if a player is eliminated after one loss and games are played until only one entrant has not lost. (Assume there are no ties.)
22. A chain letter starts when a person sends a letter to five others. Each person who receives the letter either sends it to five other people who have never received it or does not send it to anyone. Suppose that 10,000 people send out the letter before the chain ends and that no one receives more than one letter. How many people receive the letter, and how many do not send it out?
23. A chain letter starts with a person sending a letter out to 10 others. Each person is asked to send the letter out to 10 others, and each letter contains a list of the previous six people in the chain. Unless there are fewer than six names in the list, each person sends one dollar to the first person in this list, removes the name of this person from the list, moves up each of the other five names one position, and inserts his or her name at the end of this list. If no person breaks the chain and no one receives more than one letter, how much money will a person in the chain ultimately receive?

*24. Either draw a full m-ary tree with 76 leaves and height 3, where m is a positive integer, or show that no such tree exists.

*25. Either draw a full m-ary tree with 84 leaves and height 3, where m is a positive integer, or show that no such tree exists.

*26. A full m-ary tree T has 81 leaves and height 4.
   a) Give the upper and lower bounds for m.
   b) What is m if T is also balanced?

A **complete *m*-ary tree** is a full m-ary tree, where every leaf is at the same level.

27. Construct a complete binary tree of height 4 and a complete 3-ary tree of height 3.
28. How many vertices and how many leaves does a complete m-ary tree of height h have?
29. Prove
   a) part (*ii*) of Theorem 4.
   b) part (*iii*) of Theorem 4.

☞30. Show that a full m-ary balanced tree of height h has more than m^{h-1} leaves.

31. How many edges are there in a forest of t trees containing a total of n vertices?
32. Explain how a tree can be used to represent the table of contents of a book organized into chapters, where each chapter is organized into sections, and each section is organized into subsections.
33. How many different isomers do these saturated hydrocarbons have?
   a) C_3H_8 **b)** C_5H_{12} **c)** C_6H_{14}
34. What does each of these represent in an organizational tree?
   a) the parent of a vertex
   b) a child of a vertex
   c) a sibling of a vertex
   d) the ancestors of a vertex
   e) the descendants of a vertex
   f) the level of a vertex
   g) the height of the tree
35. Answer the same questions as those given in Exercise 34 for a rooted tree representing a computer file system.
36. **a)** Draw the complete binary tree with 15 vertices that represents a tree-connected network of 15 processors.

b) Show how 16 numbers can be added using the 15 processors in part (a) using four steps.

37. Let n be a power of 2. Show that n numbers can be added in $\log n$ steps using a tree-connected network of $n - 1$ processors.

***38.** A **labeled tree** is a tree where each vertex is assigned a label. Two labeled trees are considered isomorphic when there is an isomorphism between them that preserves the labels of vertices. How many nonisomorphic trees are there with three vertices labeled with different integers from the set $\{1, 2, 3\}$? How many nonisomorphic trees are there with four vertices labeled with different integers from the set $\{1, 2, 3, 4\}$?

The **eccentricity** of a vertex in an unrooted tree is the length of the longest simple path beginning at this vertex. A vertex is called a **center** if no vertex in the tree has smaller eccentricity than this vertex. In Exercises 39–41 find every vertex that is a center in the given tree.

39.

40.

41.

42. Show that a center should be chosen as the root to produce a rooted tree of minimal height from an unrooted tree.

***43.** Show that a tree has either one center or two centers that are adjacent.

44. Show that every tree can be colored using two colors.

The **rooted Fibonacci trees** T_n are defined recursively in the following way. T_1 and T_2 are both the rooted tree consisting of a single vertex, and for $n = 3, 4, \ldots$, the rooted tree T_n is constructed from a root with T_{n-1} as its left subtree and T_{n-2} as its right subtree.

45. Draw the first seven rooted Fibonacci trees.

***46.** How many vertices, leaves, and internal vertices does the rooted Fibonacci tree T_n have, where n is a positive integer? What is its height?

47. What is wrong with the following "proof" using mathematical induction of the statement that every tree with n vertices has a path of length $n - 1$. *Basis step:* Every tree with one vertex clearly has a path of length 0. *Inductive step:* Assume that a tree with n vertices has a path of length $n - 1$, which has u as its terminal vertex. Add a vertex v and the edge from u to v. The resulting tree has $n + 1$ vertices and has a path of length n. This completes the induction step.

☞***48.** Show that the average depth of a leaf in a binary tree with n vertices is $\Omega(\log n)$.

10.2 Applications of Trees

Introduction

We will discuss three problems that can be studied using trees. The first problem is: How should items in a list be stored so that an item can be easily located? The second problem is: What series of decisions should be made to find an object with a certain property in a collection of objects of a certain type? The third problem is: How should a set of characters be efficiently coded by bit strings?

Binary Search Trees

Links

Searching for items in a list is one of the most important tasks that arises in computer science. Our primary goal is to implement a searching algorithm that finds items efficiently when the

items are totally ordered. This can be accomplished through the use of a **binary search tree,** which is a binary tree in which each child of a vertex is designated as a right or left child, no vertex has more than one right child or left child, and each vertex is labeled with a key, which is one of the items. Furthermore, vertices are assigned keys so that the key of a vertex is both larger than the keys of all vertices in its left subtree and smaller than the keys of all vertices in its right subtree.

This recursive procedure is used to form the binary search tree for a list of items. Start with a tree containing just one vertex, namely, the root. The first item in the list is assigned as the key of the root. To add a new item, first compare it with the keys of vertices already in the tree, starting at the root and moving to the left if the item is less than the key of the respective vertex if this vertex has a left child, or moving to the right if the item is greater than the key of the respective vertex if this vertex has a right child. When the item is less than the respective vertex and this vertex has no left child, then a new vertex with this item as its key is inserted as a new left child. Similarly, when the item is greater than the respective vertex and this vertex has no right child, then a new vertex with this item as its key is inserted as a new right child. We illustrate this procedure with Example 1.

EXAMPLE 1 Form a binary search tree for the words *mathematics, physics, geography, zoology, meteorology, geology, psychology,* and *chemistry* (using alphabetical order).

Solution: Figure 1 displays the steps used to construct this binary search tree. The word *mathematics* is the key of the root. Because *physics* comes after *mathematics* (in alphabetical order), add a right child of the root with key *physics.* Because *geography* comes before *mathematics,* add a left child of the root with key *geography.* Next, add a right child of the vertex with key *physics,* and assign it the key *zoology,* because *zoology* comes after *mathematics* and after *physics.* Similarly, add a left child of the vertex with key *physics* and assign this new vertex the key *meteorology.* Add a right child of the vertex with key *geography* and assign this new vertex the key *geology.* Add a left child of the vertex with key *zoology* and assign it the key *psychology.*

FIGURE 1 Constructing a Binary Search Tree.

Add a left child of the vertex with key *geography* and assign it the key *chemistry*. (The reader should work through all the comparisons needed at each step.) ◀

Once we have a binary search tree, we need a way to locate items in the binary search tree, as well as a way to add new items. Algorithm 1, an insertion algorithm, actually does both of these tasks, even though it may appear that it is only designed to add vertices to a binary search tree. That is, Algorithm 1 is a procedure that locates an item x in a binary search tree if it is present, and adds a new vertex with x as its key if x is not present. In the pseudocode, v is the vertex currently under examination and $label(v)$ represents the key of this vertex. The algorithm begins by examining the root. If the x equals the key of v, then the algorithm has found the location of x and terminates; if x is less than the key of v, we move to the left child of v and repeat the procedure; and if x is greater than the key of v, we move to the right child of v and repeat the procedure. If at any step we attempt to move to a child that is not present, we know that x is not present in the tree, and we add a new vertex as this child with x as its key.

ALGORITHM 1 **Locating and Adding Items to a Binary Search Tree.**

```
procedure insertion(T: binary search tree, x: item)
v := root of T
{a vertex not present in T has the value null}
while v ≠ null and label(v) ≠ x
begin
  if x < label(v) then
    if left child of v ≠ null then v := left child of v
    else add new vertex as a left child of v and set v := null
  else
    if right child of v ≠ null then v := right child of v
    else add new vertex as a right child of v to T and set v := null
end
if root of T = null then add a vertex v to the tree and label it with x
else if v is null or label(v) ≠ x then label new vertex with x and let v be this new vertex
{v = location of x}
```

Example 2 illustrates the use of Algorithm 1 to insert a new item into a binary search tree.

EXAMPLE 2 Use Algorithm 1 to insert the word *oceanography* into the binary search tree in Example 1.

Solution: Algorithm 1 begins with v, the vertex under examination, equal to the root of T, so $label(v) = mathematics$. Because $v \neq null$ and $label(v) = mathematics < oceanography$, we next examine the right child of the root. This right child exists, so we set v, the vertex under examination, to be this right child. At this step we have $v \neq null$ and $label(v) = physics > oceanography$, so we examine the left child of v. This left child exists, so we set v, the vertex under examination, to this left child. At this step, we also have $v \neq null$ and $label(v) = metereology < oceanography$, so we try to examine the right child of v. However, this right child does not exist, so we add a new vertex as the right child of v (which at this point is the vertex with the key *metereology*) and we set $v := null$. We now exit the **while** loop because $v = null$. Because the root of T is not *null* and $v = null$, we use the **else if** statement at the end of the algorithm to label our new vertex with the key *oceanography*. ◀

FIGURE 2 Adding Unlabeled Vertices to Make a Binary Search Tree Full.

We will now determine the computational complexity of this procedure. Suppose we have a binary search tree T for a list of n items. We can form a full binary tree U from T by adding unlabeled vertices whenever necessary so that every vertex with a key has two children. This is illustrated in Figure 2. Once we have done this, we can easily locate or add a new item as a key without adding a vertex.

The most comparisons needed to add a new item is the length of the longest path in U from the root to a leaf. The internal vertices of U are the vertices of T. It follows that U has n internal vertices. We can now use part (*ii*) of Theorem 4 in Section 10.1 to conclude that U has $n + 1$ leaves. Using Corollary 1 of Section 10.1, we see that the height of U is greater than or equal to $h = \lceil \log(n + 1) \rceil$. Consequently, it is necessary to perform at least $\lceil \log(n + 1) \rceil$ comparisons to add some item. Note that if U is balanced, its height is $\lceil \log(n + 1) \rceil$ (by Corollary 1 of Section 10.1). Thus, if a binary search tree is balanced, locating or adding an item requires no more than $\lceil \log(n + 1) \rceil$ comparisons. A binary search tree can become unbalanced as items are added to it. Because balanced binary search trees give optimal worst-case complexity for binary searching, algorithms have been devised that rebalance binary search trees as items are added. The interested reader can consult references on data structures for the description of such algorithms.

Decision Trees

Links

Rooted trees can be used to model problems in which a series of decisions leads to a solution. For instance, a binary search tree can be used to locate items based on a series of comparisons, where each comparison tells us whether we have located the item, or whether we should go right or left in a subtree. A rooted tree in which each internal vertex corresponds to a decision, with a subtree at these vertices for each possible outcome of the decision, is called a **decision tree.** The possible solutions of the problem correspond to the paths to the leaves of this rooted tree. Example 3 illustrates an application of decision trees.

EXAMPLE 3 Suppose there are seven coins, all with the same weight, and a counterfeit coin that weighs less than the others. How many weighings are necessary using a balance scale to determine which of the eight coins is the counterfeit one? Give an algorithm for finding this counterfeit coin.

Extra Examples

Solution: There are three possibilities for each weighing on a balance scale. The two pans can have equal weight, the first pan can be heavier, or the second pan can be heavier. Consequently, the decision tree for the sequence of weighings is a 3-ary tree. There are at least eight leaves in

FIGURE 3 A Decision Tree for Locating a Counterfeit Coin. The counterfeit coin is shown in color below each final weighing.

the decision tree because there are eight possible outcomes (because each of the eight coins can be the counterfeit lighter coin), and each possible outcome must be represented by at least one leaf. The largest number of weighings needed to determine the counterfeit coin is the height of the decision tree. From Corollary 1 of Section 10.1 it follows that the height of the decision tree is at least $\lceil \log_3 8 \rceil = 2$. Hence, at least two weighings are needed.

It is possible to determine the counterfeit coin using two weighings. The decision tree that illustrates how this is done is shown in Figure 3. ◀

THE COMPLEXITY OF SORTING ALGORITHMS Many different sorting algorithms have been developed. To decide whether a particular sorting algorithm is efficient, its complexity is determined. Using decision trees as models, a lower bound for the worst-case complexity of sorting algorithms can be found.

We can use decision trees to model sorting algorithms and to determine an estimate for the worst-case complexity of these algorithms. Note that given n elements, there are $n!$ possible orderings of these elements, because each of the $n!$ permutations of these elements can be the correct order. The sorting algorithms studied in this book, and most commonly used sorting algorithms, are based on binary comparisons, that is, the comparison of two elements at a time. The result of each such comparison narrows down the set of possible orderings. Thus, a sorting algorithm based on binary comparisons can be represented by a binary decision tree in which each internal vertex represents a comparison of two elements. Each leaf represents one of the $n!$ permutations of n elements.

EXAMPLE 4 We display in Figure 4 a decision tree that orders the elements of the list a, b, c. ◀

The complexity of a sort based on binary comparisons is measured in terms of the number of such comparisons used. The largest number of binary comparisons ever needed to sort a list with n elements gives the worst-case performance of the algorithm. The most comparisons used equals the longest path length in the decision tree representing the sorting procedure. In other words, the largest number of comparisons ever needed is equal to the height of the decision tree. Because the height of a binary tree with $n!$ leaves is at least $\lceil \log n! \rceil$ (using Corollary 1 in Section 10.1), at least $\lceil \log n! \rceil$ comparisons are needed, as stated in Theorem 1.

THEOREM 1 A sorting algorithm based on binary comparisons requires at least $\lceil \log n! \rceil$ comparisons.

FIGURE 4 A Decision Tree for Sorting Three Distinct Elements.

We can use Theorem 1 to provide a big-Omega estimate for the number of comparisons used by a sorting algorithm based on binary comparison. We need only note that by Exercise 62 in Section 3.2 we know that $\lceil \log n! \rceil$ is $\Theta(n \log n)$, one of the commonly used reference functions for the computational complexity of algorithms. Corollary 1 is a consequence of this estimate.

COROLLARY 1 The number of comparisons used by a sorting algorithm to sort n elements based on binary comparisons is $\Omega(n \log n)$.

A consequence of Corollary 1 is that a sorting algorithm based on binary comparisons that uses $\Theta(n \log n)$ comparisons, in the worst case, to sort n elements is optimal, in the sense that no other such algorithm has better worst-case complexity. Note that by Theorem 1 in Section 4.4 we see that the merge sort algorithm is optimal in this sense.

We can also establish a similar result for the average-case complexity of sorting algorithms. The average number of comparisons used by a sorting algorithm based on binary comparisons is the average depth of a leaf in the decision tree representing the sorting algorithm. By Exercise 48 in Section 10.1 we know that the average depth of a leaf in a binary tree with N vertices is $\Omega(\log N)$. We obtain the following estimate when we let $N = n!$ and note that a function that is $\Omega(\log n!)$ is also $\Omega(n \log n)$ because $\log n!$ is $\Theta(n \log n)$.

THEOREM 2 The average number of comparisons used by a sorting algorithm to sort n elements based on binary comparisons is $\Omega(n \log n)$.

Prefix Codes

Consider the problem of using bit strings to encode the letters of the English alphabet (where no distinction is made between lowercase and uppercase letters). We can represent each letter with a bit string of length five, because there are only 26 letters and there are 32 bit strings of length five. The total number of bits used to encode data is five times the number of characters in the text when each character is encoded with five bits. Is it possible to find a coding scheme

of these letters such that, when data are coded, fewer bits are used? We can save memory and reduce transmittal time if this can be done.

Consider using bit strings of different lengths to encode letters. Letters that occur more frequently should be encoded using short bit strings, and longer bit strings should be used to encode rarely occurring letters. When letters are encoded using varying numbers of bits, some method must be used to determine where the bits for each character start and end. For instance, if e were encoded with 0, a with 1, and t with 01, then the bit string 0101 could correspond to *eat, tea, eaea,* or *tt.*

FIGURE 5 The Binary Tree with a Prefix Code.

One way to ensure that no bit string corresponds to more than one sequence of letters is to encode letters so that the bit string for a letter never occurs as the first part of the bit string for another letter. Codes with this property are called **prefix codes.** For instance, the encoding of e as 0, a as 10, and t as 11 is a prefix code. A word can be recovered from the unique bit string that encodes its letters. For example, the string 10110 is the encoding of *ate.* To see this, note that the initial 1 does not represent a character, but 10 does represent a (and could not be the first part of the bit string of another letter). Then, the next 1 does not represent a character, but 11 does represent t. The final bit, 0, represents e.

A prefix code can be represented using a binary tree, where the characters are the labels of the leaves in the tree. The edges of the tree are labeled so that an edge leading to a left child is assigned a 0 and an edge leading to a right child is assigned a 1. The bit string used to encode a character is the sequence of labels of the edges in the unique path from the root to the leaf that has this character as its label. For instance, the tree in Figure 5 represents the encoding of e by 0, a by 10, t by 110, n by 1110, and s by 1111.

The tree representing a code can be used to decode a bit string. For instance, consider the word encoded by 11111011100 using the code in Figure 5. This bit string can be decoded by starting at the root, using the sequence of bits to form a path that stops when a leaf is reached. Each 0 bit takes the path down the edge leading to the left child of the last vertex in the path, and each 1 bit corresponds to the right child of this vertex. Consequently, the initial 1111 corresponds to the path starting at the root, going right four times, leading to a leaf in the graph that has s as its label, because the string 1111 is the code for s. Continuing with the fifth bit, we reach a leaf next after going right then left, when the vertex labeled with a, which is encoded by 10, is visited. Starting with the seventh bit, we reach a leaf next after going right three times and then left, when the vertex labeled with n, which is encoded by 1110, is visited. Finally, the last bit, 0, leads to the leaf that is labeled with e. Therefore, the original word is *sane.*

We can construct a prefix code from any binary tree where the left edge at each internal vertex is labeled by 0 and the right edge by a 1 and where the leaves are labeled by characters. Characters are encoded with the bit string constructed using the labels of the edges in the unique path from the root to the leaves.

Links

HUFFMAN CODING We now introduce an algorithm that takes as input the frequencies (which are the probabilities of occurrences) of symbols in a string and produces as output a prefix code that encodes the string using the fewest possible bits, among all possible binary prefix codes for these symbols. This algorithm, known as **Huffman coding,** was developed by David Huffman in a term paper he wrote in 1951 while a graduate student at MIT. (Note that this algorithm assumes that we already know how many times each symbol occurs in the string, so we can compute the frequency of each symbol by dividing the number of times this symbol occurs by the length of the string.) Huffman coding is a fundamental algorithm in *data compression,* the subject devoted to reducing the number of bits required to represent information. Huffman coding is extensively used to compress bit strings representing text and it also plays an important role in compressing audio and image files.

Algorithm 2 presents the Huffman coding algorithm. Given symbols and their frequencies, our goal is to construct a rooted binary tree where the symbols are the labels of the leaves. The algorithm begins with a forest of trees each consisting of one vertex, where each vertex has a symbol as its label and where the weight of this vertex equals the frequency of the symbol that

is its label. At each step, we combine two trees having the least total weight into a single tree by introducing a new root and placing the tree with larger weight as its left subtree and the tree with smaller weight as its right subtree. Furthermore, we assign the sum of the weights of the two subtrees of this tree as the total weight of the tree.(Although procedures for breaking ties by choosing between trees with equal weights can be specified, we will not specify such procedures here.) The algorithm is finished when it has constructed a tree, that is, when the forest is reduced to a single tree.

Demo

ALGORITHM 2 Huffman Coding.

procedure *Huffman*(C: symbols a_i with frequencies w_i, $i = 1, \ldots, n$)
$F :=$ forest of n rooted trees, each consisting of the single vertex a_i and assigned weight w_i
while F is not a tree
begin
  Replace the rooted trees T and T' of least weights from F with $w(T) \geq w(T')$ with a tree having a new root that has T as its left subtree and T' as its right subtree. Label the new edge to T with 0 and the new edge to T' with 1.
  Assign $w(T) + w(T')$ as the weight of the new tree.
end
{the Huffman coding for the symbol a_i is the concatenation of the labels of the edges in the unique path from the root to the vertex a_i}

Example 5 illustrates how Algorithm 2 is used to encode a set of five symbols.

EXAMPLE 5 Use Huffman coding to encode the following symbols with the frequencies listed: A: 0.08, B: 0.10, C: 0.12, D: 0.15, E: 0.20, F: 0.35. What is the average number of bits used to encode a character?

Extra Examples

Solution: Figure 6 displays the steps used to encode these symbols. The encoding produced encodes A by 111, B by 110, C by 011, D by 010, E by 10, and F by 00. The average number of bits used to encode a symbol using this encoding is $3 \cdot 0.08 + 3 \cdot 0.10 + 3 \cdot 0.12 + 3 \cdot 0.15 + 2 \cdot 0.20 + 2 \cdot 0.35 = 2.45$. ◀

Links

DAVID A. HUFFMAN (1925–1999) David Huffman grew up in Ohio. At the age of 18 he received his B.S. in electrical engineering from The Ohio State University. Afterward he served in the U.S. Navy as a radar maintenance officer on a destroyer that had the mission of clearing mines in Asian waters after World War II. Later, he earned his M.S. from Ohio State and his Ph.D. in electrical engineering from MIT. Huffman joined the MIT faculty in 1953, where he remained until 1967 when he became the founding member of the computer science department at the University of California at Santa Cruz. He played an important role in developing this department and spent the remainder of his career there, retiring in 1994.

Huffman is noted for his contributions to information theory and coding, signal designs for radar and for communications, and design procedures for asynchronous logical circuits. His work on surfaces with zero curvature led him to develop original techniques for folding paper and vinyl into unusual shapes considered works of art by many and publicly displayed in several exhibits. However, Huffman is best known for his development of what is now called Huffman coding, a result of a term paper he wrote during his graduate work at MIT.

Huffman enjoyed exploring the outdoors, hiking, and traveling extensively. He became certified as a scuba diver when he was in his late 60s. He kept poisonous snakes as pets.

FIGURE 6 Huffman Coding of Symbols in Example 4.

Note that Huffman coding is a greedy algorithm. Replacing the two subtrees with the smallest weight at each step leads to an optimal code in the sense that no binary prefix code for these symbols can encode these symbols using fewer bits. We leave the proof that Huffman codes are optimal as Exercise 32 at the end of this section.

There are many variations of Huffman coding. For example, instead of encoding single symbols, we can encode blocks of symbols of a specified length, such as blocks of two symbols. Doing so may reduce the number of bits required to encode the string (see Exercise 30 at the end of this section). We can also use more than two symbols to encode the original symbols in the string (see the preamble to Exercise 28 at the end of this section). Furthermore, a variation known as adaptive Huffman coding (see [Sa00]) can be used when the frequency of each symbol in a string is not known in advance, so that encoding is done at the same time the string is being read.

Game Trees

Links

Trees can be used to analyze certain types of games such as tic-tac-toe, nim, checkers, and chess. In each of these games, two players take turns making moves. Each player knows the moves made by the other player and no element of chance enters into the game. We model such games using **game trees;** the vertices of these trees represent the positions that a game can be in as it progresses; the edges represent legal moves between these positions. Because game trees are usually large, we simplify game trees by representing all symmetric positions of a game by the same vertex. However, the same position of a game may be represented by different vertices if different sequences of moves lead to this position. The root represents the starting position. The usual convention is to represent vertices at even levels by boxes and vertices at odd levels by circles. When the game is in a position represented by a vertex at an even level, it is the first player's move; when the game is in a position represented by a vertex at an odd level, it is the second player's move. Game trees may be infinite when the games they represent never end, such as games that can enter infinite loops, but for most games there are rules that lead to finite game trees.

The leaves of a game tree represent the final positions of a game. We assign a value to each leaf indicating the payoff to the first player if the game terminates in the position represented by this leaf. For games that are win–lose, we label a terminal vertex represented by a circle with a 1 to indicate a win by the first player and we label a terminal vertex represented by a box with a -1 to indicate a win by the second player. For games where draws are allowed, we label a terminal vertex corresponding to a draw position with a 0. Note that for win–lose games, we have assigned values to terminal vertices so that the larger the value, the better the outcome for the first player.

In Example 6 we display a game tree for a well-known and well-studied game.

EXAMPLE 6 **Nim** In a version of the game of **nim,** at the start of a game there are a number of piles of stones. Two players take turns making moves; a legal move consists of removing one or more stones from one of the piles, without removing all the stones left. A player without a legal move loses. (Another way to look at this is that the player removing the last stone loses because the position with no piles of stones is not allowed.) The game tree shown in Figure 7 represents

FIGURE 7 The Game Tree for a Game of Nim.

FIGURE 8 Some of the Game Tree for Tic-Tac-Toe.

this version of nim given the starting position where there are three piles of stones containing two, two, and one stone each, respectively. We represent each position with an unordered list of the number of stones in the different piles (the order of the piles does not matter). The initial move by the first player can lead to three possible positions because this player can remove one stone from a pile with two stones (leaving three piles containing one, one, and two stones); two stones from a pile containing two stones (leaving two piles containing two stones and one stone); or one stone from the pile containing one stone (leaving two piles of two stones). When only one pile with one stone is left, no legal moves are possible, so such positions are terminal positions. Because nim is a win–lose game, we label the terminal vertices with $+1$ when they represent wins for the first player and -1 when they represent wins for the second player. ◀

EXAMPLE 7 **Tic-tac-toe** The game tree for tic-tac-toe is extremely large and cannot be drawn here, although a computer could easily build such a tree. We show a portion of the game tic-tac-toe in Figure 8(a). Note that by considering symmetric positions equivalent, we need only consider three possible initial moves, as shown in Figure 8(a). We also show a subtree of this game tree leading to terminal positions in Figure 8(b), where a player who can win makes a winning move. ◀

We can recursively define the values of all vertices in a game tree in a way that enables us to determine the outcome of this game when both players follow optimal strategies. By a **strategy** we mean a set of rules that tells a player how to select moves to win the game. An optimal strategy for the first player is a strategy that maximizes the payoff to this player and for the second player is a strategy that minimizes this payoff. We now recursively define the value of a vertex.

DEFINITION 1 The *value of a vertex in a game tree* is defined recursively as:

(*i*) the value of a leaf is the payoff to the first player when the game terminates in the position represented by this leaf.

(*ii*) the value of an internal vertex at an even level is the maximum of the values of its children, and the value of an internal vertex at an odd level is the minimum of the values of its children.

The strategy where the first player moves to a position represented by a child with maximum value and the second player moves to a position of a child with minimum value is called the **minmax strategy.** We can determine who will win the game when both players follow the minmax strategy by calculating the value of the root of the tree; this value is called the **value** of the tree. This is a consequence of Theorem 3.

THEOREM 3 The value of a vertex of a game tree tells us the payoff to the first player if both players follow the minmax strategy and play starts from the position represented by this vertex.

Proof: We will use induction to prove this theorem.

BASIS STEP: If the vertex is a leaf, by definition the value assigned to this vertex is the payoff to the first player.

INDUCTIVE STEP: The inductive hypothesis is the assumption that the values of the children of a vertex are the payoffs to the first player, assuming that play starts at each of the positions represented by these vertices. We need to consider two cases, when it is the first player's turn and when it is the second player's turn.

When it is the first player's turn, this player follows the minmax strategy and moves to the position represented by the child with the largest value. By the inductive hypothesis, this value is the payoff to the first player when play starts at the position represented by this child and follows the minmax strategy. By the recursive step in the definition of the value of an internal vertex at an even level (as the maximum value of its children), the value of this vertex is the payoff when play begins at the position represented by this vertex.

When it is the second player's turn, this player follows the minmax strategy and moves to the position represented by the child with the least value. By the inductive hypothesis, this value is the payoff to the first player when play starts at the position represented by this child and both players follow the minmax strategy. By the recursive definition of the value of an internal vertex at an odd level as the minimum value of its children, the value of this vertex is the payoff when play begins at the position represented by this vertex. ◁

Remark: By extending the proof of Theorem 3, it can be shown that the minmax strategy is the optimal strategy for both players.

Example 8 illustrates how the minmax procedure works. It displays the values assigned to the internal vertices in the game tree from Example 6. Note that we can shorten the computation required by noting that for win–lose games, once a child of a square vertex with value $+1$ is found, the value of the square vertex is also $+1$ because $+1$ is the largest possible payoff.

FIGURE 9 Showing the Values of Vertices in the Game of Nim.

Similarly, once a child of a circle vertex with value -1 is found, this is the value of the circle vertex also.

EXAMPLE 8 In Example 6 we constructed the game tree for nim with a starting position where there are three piles containing two, two, and one stones. In Figure 9 we show the values of the vertices of this game tree. The values of the vertices are computed using the values of the leaves and working one level up at a time. In the right margin of this figure we indicate whether we use the maximum or minimum of the values of the children to find the value of an internal vertex at each level. For example, once we have found the values of the three children of the root, which are 1, -1, and -1, we find the value of the root by computing $\max(1, -1, -1) = 1$. Because the value of the root is 1, it follows that the first player wins when both players follow a minmax strategy. ◀

Game trees for some well-known games can be extraordinarily large, because these games have many different possible moves. For example, the game tree for chess has been estimated to have as many as 10^{100} vertices! It may be impossible to use Theorem 3 directly to study a game because of the size of the game tree. Therefore, various approaches have been devised to help determine good strategies and to determine the outcome of such games. One useful technique, called *alpha-beta pruning,* eliminates much computation by pruning portions of the game tree that cannot affect the values of ancestor vertices. (For information about alpha-beta pruning, consult [Gr90].) Another useful approach is to use *evaluation functions,* which estimate the value of internal vertices in the game tree when it is not feasible to compute these values exactly. For example, in the game of tic-tac-toe, as an evaluation function for a position, we may use the number of files (rows, columns, and diagonals) containing no Os (used to indicate moves of the second player) minus the number of files containing no Xs (used to indicate moves of the first player). This evaluation function provides some indication of which player has the advantage in the game. Once the values of an evaluation function are inserted, the value of the game can be computed following the rules used for the minmax strategy. Computer programs created to play chess, such as the famous Deep Blue program, are based on sophisticated evaluation functions. For more information about how computers play chess see [Le91].

Links

Exercises

1. Build a binary search tree for the words *banana, peach, apple, pear, coconut, mango,* and *papaya* using alphabetical order.
2. Build a binary search tree for the words *oenology, phrenology, campanology, ornithology, ichthyology, limnology, alchemy,* and *astrology* using alphabetical order.
3. How many comparisons are needed to locate or to add each of these words in the search tree for Exercise 1, starting fresh each time?

   a) *pear* **b)** *banana*
   c) *kumquat* **d)** *orange*
4. How many comparisons are needed to locate or to add each of the words in the search tree for Exercise 2, starting fresh each time?

   a) *palmistry* **b)** *etymology*
   c) *paleontology* **d)** *glaciology*
5. Using alphabetical order, construct a binary search tree for the words in the sentence "*The quick brown fox jumps over the lazy dog.*"
6. How many weighings of a balance scale are needed to find a lighter counterfeit coin among four coins? Describe an algorithm to find the lighter coin using this number of weighings.
7. How many weighings of a balance scale are needed to find a counterfeit coin among four coins if the counterfeit coin may be either heavier or lighter than the others? Describe an algorithm to find the counterfeit coin using this number of weighings.

*8. How many weighings of a balance scale are needed to find a counterfeit coin among eight coins if the counterfeit coin is either heavier or lighter than the others? Describe an algorithm to find the counterfeit coin using this number of weighings.

*9. How many weighings of a balance scale are needed to find a counterfeit coin among 12 coins if the counterfeit coin is lighter than the others? Describe an algorithm to find the lighter coin using this number of weighings.

*10. One of four coins may be counterfeit. If it is counterfeit, it may be lighter or heavier than the others. How many weighings are needed, using a balance scale, to determine whether there is a counterfeit coin, and if there is, whether it is lighter or heavier than the others? Describe an algorithm to find the counterfeit coin and determine whether it is lighter or heavier using this number of weighings.

11. Find the least number of comparisons needed to sort four elements and devise an algorithm that sorts these elements using this number of comparisons.

*12. Find the least number of comparisons needed to sort five elements and devise an algorithm that sorts these elements using this number of comparisons.

The **tournament sort** is a sorting algorithm that works by building an ordered binary tree. We represent the elements to be sorted by vertices that will become the leaves. We build up the tree one level at a time as we would construct the tree representing the winners of matches in a tournament. Working left to right, we compare pairs of consecutive elements, adding a parent vertex labeled with the larger of the two elements under comparison. We make similar comparisons between labels of vertices at each level until we reach the root of the tree that is labeled with the largest element. The tree constructed by the tournament sort of 22, 8, 14, 17, 3, 9, 27, 11 is illustrated in part (a) of the figure. Once the largest element has been determined, the leaf with this label is relabeled by $-\infty$, which is defined to be less than every element. The labels of all vertices on the path from this vertex up to the root of the tree are recalculated, as shown in part (b) of the figure. This produces the second largest element. This process continues until the entire list has been sorted.

13. Complete the tournament sort of the list 22, 8, 14, 17, 3, 9, 27, 11. Show the labels of the vertices at each step.
14. Use the tournament sort to sort the list 17, 4, 1, 5, 13, 10, 14, 6.
15. Describe the tournament sort using pseudocode.
16. Assuming that n, the number of elements to be sorted, equals 2^k for some positive integer k, determine the number of comparisons used by the tournament sort to find the largest element of the list using the tournament sort.
17. How many comparisons does the tournament sort use to find the second largest, the third largest, and so on, up to the $(n-1)$st largest (or second smallest) element?
18. Show that the tournament sort requires $\Theta(n \log n)$ comparisons to sort a list of n elements. [*Hint:* By inserting the appropriate number of dummy elements defined to be

smaller than all integers, such as $-\infty$, assume that $n = 2^k$ for some positive integer k.]

19. Which of these codes are prefix codes?

a) a: 11, e: 00, t: 10, s: 01
b) a: 0, e: 1, t: 01, s: 001
c) a: 101, e: 11, t: 001, s: 011, n: 010
d) a: 010, e: 11, t: 011, s: 1011, n: 1001, i: 10101

20. Construct the binary tree with prefix codes representing these coding schemes.

a) a: 11, e: 0, t: 101, s: 100
b) a: 1, e: 01, t: 001, s: 0001, n: 00001
c) a: 1010, e: 0, t: 11, s: 1011, n: 1001, i: 10001

21. What are the codes for a, e, i, k, o, p, and u if the coding scheme is represented by this tree?

22. Given the coding scheme a: 001, b: 0001, e: 1, r: 0000, s: 0100, t: 011, x: 01010, find the word represented by

a) 01110100011. **b)** 0001110000.
c) 0100101010. **d)** 01100101010.

23. Use Huffman coding to encode these symbols with given frequencies: a: 0.20, b: 0.10, c: 0.15, d: 0.25, e: 0.30. What is the average number of bits required to encode a character?

24. Use Huffman coding to encode these symbols with given frequencies: A: 0.10, B: 0.25, C: 0.05, D: 0.15, E: 0.30, F: 0.07, G: 0.08. What is the average number of bits required to encode a symbol?

25. Construct two different Huffman codes for these symbols and frequencies: t: 0.2, u: 0.3, v: 0.2, w: 0.3.

26. **a)** Use Huffman coding to encode these symbols with frequencies a: 0.4, b: 0.2, c: 0.2, d: 0.1, e: 0.1 in two different ways by breaking ties in the algorithm differently. First, among the trees of minimum weight select two trees with the largest number of vertices to combine at each stage of the algorithm. Second, among the trees of minimum weight select two trees with the smallest number of vertices at each stage.

b) Compute the average number of bits required to encode a symbol with each code and compute the variances of this number of bits for each code. Which tie-breaking procedure produced the smaller variance in the number of bits required to encode a symbol?

27. Construct a Huffman code for the letters of the English alphabet where the frequencies of letters in typical English text are as shown in this table.

Letter	*Frequency*	*Letter*	*Frequency*
A	0.0817	N	0.0662
B	0.0145	O	0.0781
C	0.0248	P	0.0156
D	0.0431	Q	0.0009
E	0.1232	R	0.0572
F	0.0209	S	0.0628
G	0.0182	T	0.0905
H	0.0668	U	0.0304
I	0.0689	V	0.0102
J	0.0010	W	0.0264
K	0.0080	X	0.0015
L	0.0397	Y	0.0211
M	0.0277	Z	0.0005

Suppose that m is a positive integer with $m \geq 2$. An m-ary Huffman code for a set of N symbols can be constructed analogously to the construction of a binary Huffman code. At the initial step, $((N-1)$ **mod** $(m-1)) + 1$ trees consisting of a single vertex with least weights are combined into a rooted tree with these vertices as leaves. At each subsequent step, the m trees of least weight are combined into an m-ary tree.

28. Describe the m-ary Huffman coding algorithm in pseudocode.

29. Using the symbols 0, 1, and 2 use ternary ($m = 3$) Huffman coding to encode these letters with the given frequencies: A: 0.25, E: 0.30, N: 0.10, R: 0.05, T: 0.12, Z: 0.18.

30. Consider the three symbols A, B, and C with frequencies A: 0.80, B: 0.19, C: 0.01.

a) Construct a Huffman code for these three symbols.
b) Form a new set of nine symbols by grouping together blocks of two symbols, AA, AB, AC, BA, BB, BC, CA, CB, and CC. Construct a Huffman code for these nine symbols, assuming that the occurrences of symbols in the original text are independent.
c) Compare the average number of bits required to encode text using the Huffman code for the three symbols in part (a) and the Huffman code for the nine blocks of two symbols constructed in part (b). Which is more efficient?

31. Given $n+1$ symbols $x_1, x_2, \ldots, x_n, x_{n+1}$ appearing 1, $f_1, f_2, \ldots, f_n$ times in a symbol string, respectively, where f_j is the jth Fibonacci number, what is the maximum number of bits used to encode a symbol when all possible tie-breaking selections are considered at each stage of the Huffman coding algorithm?

‘32. Show that Huffman codes are optimal in the sense that they represent a string of symbols using the fewest bits among all binary prefix codes.

33. Draw a game tree for nim if the starting position consists of two piles with two and three stones, respectively. When drawing the tree represent by the same vertex symmetric positions that result from the same move. Find the value of each vertex of the game tree. Who wins the game if both players follow an optimal strategy?

34. Draw a game tree for nim if the starting position consists of three piles with one, two, and three stones, respectively. When drawing the tree represent by the same vertex symmetric positions that result from the same move. Find the value of each vertex of the game tree. Who wins the game if both players follow an optimal strategy?

35. Suppose that we vary the payoff to the winning player in the game of nim so that the payoff is n dollars when n is the number of legal moves made before a terminal position is reached. Find the payoff to the first player if the initial position consists of
   a) two piles with one and three stones, respectively.
   b) two piles with two and four stones, respectively.
   c) three piles with one, two, and three stones, respectively.

36. Suppose that in a variation of the game of nim we allow a player to either remove one or more stones from a pile or merge the stones from two piles into one pile as long as at least one stone remains. Draw the game tree for this variation of nim if the starting position consists of three piles containing two, two, and one stone, respectively. Find the values of each vertex in the game tree and determine the winner if both players follow an optimal strategy.

37. Draw the subtree of the game tree for tic-tac-toe beginning at each of these positions. Determine the value of each of these subtrees.

a)

O	X	X
X	O	O
		X

b)

X	O	X
O	X	X
		O

c)

X		O
O	O	
X		X

d)

	O	X
	X	O
X	O	

38. Suppose that the first four moves of a tic-tac-toe game are as shown. Does the first player (whose moves are marked by Xs) have a strategy that will always win?

a)

X	O	X
	O	

b)

X	O	X
	O	

c)

	O	X
	X	
O		

d)

X		
X	O	
O		

39. Show that if a game of nim begins with two piles containing the same number of stones, as long as this number is at least two, then the second player wins when both players follow optimal strategies.

40. Show that if a game of nim begins with two piles containing different numbers of stones, the first player wins when both players follow optimal strategies.

41. How many children does the root of the game tree for checkers have? How many grandchildren does it have?

42. How many children does the root of the game tree for nim have and how many grandchildren does it have if the starting position is
   a) piles with four and five stones, respectively.
   b) piles with two, three, and four stones, respectively.
   c) piles with one, two, three, and four stones, respectively.
   d) piles with two, two, three, three, and five stones, respectively.

43. Draw the game tree for the game of tic-tac-toe for the levels corresponding to the first two moves. Assign the value of the evaluation function mentioned in the text that assigns to a position the number of files containing no Os minus the number of files containing no Xs as the value of each vertex at this level and compute the value of the tree for vertices as if the evaluation function gave the correct values for these vertices.

44. Use pseudocode to describe an algorithm for determining the value of a game tree when both players follow a minmax strategy.

10.3 Tree Traversal

Introduction

Links

Ordered rooted trees are often used to store information. We need procedures for visiting each vertex of an ordered rooted tree to access data. We will describe several important algorithms for visiting all the vertices of an ordered rooted tree. Ordered rooted trees can also be used to represent various types of expressions, such as arithmetic expressions involving numbers,

variables, and operations. The different listings of the vertices of ordered rooted trees used to represent expressions are useful in the evaluation of these expressions.

Universal Address Systems

Procedures for traversing all vertices of an ordered rooted tree rely on the orderings of children. In ordered rooted trees, the children of an internal vertex are shown from left to right in the drawings representing these directed graphs.

We will describe one way we can totally order the vertices of an ordered rooted tree. To produce this ordering, we must first label all the vertices. We do this recursively:

1. Label the root with the integer 0. Then label its k children (at level 1) from left to right with $1, 2, 3, \ldots, k$.
2. For each vertex v at level n with label A, label its k_v children, as they are drawn from left to right, with $A.1, A.2, \ldots, A.k_v$.

Following this procedure, a vertex v at level n, for $n \geq 1$, is labeled $x_1.x_2.\ldots.x_n$, where the unique path from the root to v goes through the x_1st vertex at level 1, the x_2nd vertex at level 2, and so on. This labeling is called the **universal address system** of the ordered rooted tree.

We can totally order the vertices using the lexicographic ordering of their labels in the universal address system. The vertex labeled $x_1.x_2.\ldots.x_n$ is less than the vertex labeled $y_1.y_2.\ldots.y_m$ if there is an i, $0 \leq i \leq n$, with $x_1 = y_1, x_2 = y_2, \ldots, x_{i-1} = y_{i-1}$, and $x_i < y_i$; or if $n < m$ and $x_i = y_i$ for $i = 1, 2, \ldots, n$.

EXAMPLE 1 We display the labelings of the universal address system next to the vertices in the ordered rooted tree shown in Figure 1. The lexicographic ordering of the labelings is

Extra Examples

$$0 < 1 < 1.1 < 1.2 < 1.3 < 2 < 3 < 3.1 < 3.1.1 < 3.1.2 < 3.1.2.1 < 3.1.2.2 < 3.1.2.3 < 3.1.2.4 < 3.1.3 < 3.2 < 4 < 4.1 < 5 < 5.1 < 5.1.1 < 5.2 < 5.3$$ ◀

FIGURE 1 The Universal Address System of an Ordered Rooted Tree.

FIGURE 2 Preorder Traversal.

FIGURE 3 The Ordered Rooted Tree *T*.

Traversal Algorithms

Procedures for systematically visiting every vertex of an ordered rooted tree are called **traversal algorithms.** We will describe three of the most commonly used such algorithms, **preorder traversal, inorder traversal,** and **postorder traversal.** Each of these algorithms can be defined recursively. We first define preorder traversal.

DEFINITION 1 Let T be an ordered rooted tree with root r. If T consists only of r, then r is the *preorder traversal* of T. Otherwise, suppose that $T_1, T_2, \ldots, T_n$ are the subtrees at r from left to right in T. The *preorder traversal* begins by visiting r. It continues by traversing T_1 in preorder, then T_2 in preorder, and so on, until T_n is traversed in preorder.

The reader should verify that the preorder traversal of an ordered rooted tree gives the same ordering of the vertices as the ordering obtained using a universal address system. Figure 2 indicates how a preorder traversal is carried out.

Example 2 illustrates preorder traversal.

EXAMPLE 2 In which order does a preorder traversal visit the vertices in the ordered rooted tree T shown in Figure 3?

Extra Examples

Solution: The steps of the preorder traversal of T are shown in Figure 4. We traverse T in preorder by first listing the root a, followed by the preorder list of the subtree with root b, the preorder list of the subtree with root c (which is just c) and the preorder list of the subtree with root d.

The preorder list of the subtree with root b begins by listing b, then the vertices of the subtree with root e in preorder, and then the subtree with root f in preorder (which is just f). The preorder list of the subtree with root d begins by listing d, followed by the preorder list of the subtree with root g, followed by the subtree with root h (which is just h), followed by the subtree with root i (which is just i).

FIGURE 4 The Preorder Traversal of T.

The preorder list of the subtree with root e begins by listing e, followed by the preorder listing of the subtree with root j (which is just j), followed by the preorder listing of the subtree with root k. The preorder listing of the subtree with root g is g followed by l, followed by m. The preorder listing of the subtree with root k is k, n, o, p. Consequently, the preorder traversal of T is $a, b, e, j, k, n, o, p, f, c, d, g, l, m, h, i$. ◀

We will now define inorder traversal.

FIGURE 5 Inorder Traversal.

DEFINITION 2 Let T be an ordered rooted tree with root r. If T consists only of r, then r is the *inorder traversal* of T. Otherwise, suppose that $T_1, T_2, \ldots, T_n$ are the subtrees at r from left to right. The *inorder traversal* begins by traversing T_1 in inorder, then visiting r. It continues by traversing T_2 in inorder, then T_3 in inorder, ..., and finally T_n in inorder.

Figure 5 indicates how inorder traversal is carried out. Example 3 illustrates how inorder traversal is carried out for a particular tree.

EXAMPLE 3 In which order does an inorder traversal visit the vertices of the ordered rooted tree T in Figure 3?

Extra Examples

Solution: The steps of the inorder traversal of the ordered rooted tree T are shown in Figure 6. The inorder traversal begins with an inorder traversal of the subtree with root b, the root a, the inorder listing of the subtree with root c, which is just c, and the inorder listing of the subtree with root d.

The inorder listing of the subtree with root b begins with the inorder listing of the subtree with root e, the root b, and f. The inorder listing of the subtree with root d begins with the inorder listing of the subtree with root g, followed by the root d, followed by h, followed by i.

The inorder listing of the subtree with root e is j, followed by the root e, followed by the inorder listing of the subtree with root k. The inorder listing of the subtree with root g is l, g, m. The inorder listing of the subtree with root k is n, k, o, p. Consequently, the inorder listing of the ordered rooted tree is $j, e, n, k, o, p, b, f, a, c, l, g, m, d, h, i$. ◀

We now define postorder traversal.

DEFINITION 3 Let T be an ordered rooted tree with root r. If T consists only of r, then r is the *postorder traversal* of T. Otherwise, suppose that $T_1, T_2, \ldots, T_n$ are the subtrees at r from left to right. The *postorder traversal* begins by traversing T_1 in postorder, then T_2 in postorder, ..., then T_n in postorder, and ends by visiting r.

Figure 7 illustrates how postorder traversal is done. Example 4 illustrates how postorder traversal works.

FIGURE 6 The Inorder Traversal of *T*.

EXAMPLE 4 In which order does a postorder traversal visit the vertices of the ordered rooted tree T shown in Figure 3?

Extra Examples

Solution: The steps of the postorder traversal of the ordered rooted tree T are shown in Figure 8. The postorder traversal begins with the postorder traversal of the subtree with root b, the postorder traversal of the subtree with root c, which is just c, the postorder traversal of the subtree with root d, followed by the root a.

FIGURE 7 Postorder Traversal.

The postorder traversal of the subtree with root b begins with the postorder traversal of the subtree with root e, followed by f, followed by the root b. The postorder traversal of the rooted tree with root d begins with the postorder traversal of the subtree with root g, followed by h, followed by i, followed by the root d.

The postorder traversal of the subtree with root e begins with j, followed by the postorder traversal of the subtree with root k, followed by the root e. The postorder traversal of the subtree with root g is l, m, g. The postorder traversal of the subtree with root k is n, o, p, k. Therefore, the postorder traversal of T is $j, n, o, p, k, e, f, b, c, l, m, g, h, i, d, a$. ◀

There are easy ways to list the vertices of an ordered rooted tree in preorder, inorder, and postorder. To do this, first draw a curve around the ordered rooted tree starting at the root, moving along the edges, as shown in the example in Figure 9. We can list the vertices in preorder by listing each vertex the first time this curve passes it. We can list the vertices in inorder by listing a leaf the first time the curve passes it and listing each internal vertex the second time the curve passes it. We can list the vertices in postorder by listing a vertex the last time it is passed on the way back up to its parent. When this is done in the rooted tree in Figure 9, it follows that the preorder traversal gives $a, b, d, h, e, i, j, c, f, g, k$, the inorder traversal gives $h, d, b, i, e, j, a, f, c, k, g$; and the postorder traversal gives $h, d, i, j, e, b, f, k, g, c, a$.

Algorithms for traversing ordered rooted trees in preorder, inorder, or postorder are most easily expressed recursively.

ALGORITHM 1 **Preorder Traversal.**

```
procedure preorder(T: ordered rooted tree)
r := root of T
list r
for each child c of r from left to right
begin
  T(c) := subtree with c as its root
  preorder(T(c))
end
```

FIGURE 8 The Postorder Traversal of *T*.

FIGURE 9 A Shortcut for Traversing an Ordered Rooted Tree in Preorder, Inorder, and Postorder.

ALGORITHM 2 **Inorder Traversal.**

```
procedure inorder(T: ordered rooted tree)
r := root of T
if r is a leaf then list r
else
begin
  l := first child of r from left to right
  T(l) := subtree with l as its root
  inorder(T(l))
  list r
  for each child c of r except for l from left to right
    T(c) := subtree with c as its root
    inorder(T(c))
end
```

ALGORITHM 3 **Postorder Traversal.**

```
procedure postorder(T: ordered rooted tree)
r := root of T
for each child c of r from left to right
begin
  T(c) := subtree with c as its root
  postorder(T(c))
end
list r
```

Note that both the preorder traversal and the postorder traversal encode the structure of an ordered rooted tree when the number of children of each vertex is specified. That is, an ordered rooted tree is uniquely determined when we specify a list of vertices generated by a preorder traversal or by a postorder traversal of the tree, together with the number of children of each vertex

FIGURE 10 A Binary Tree Representing $((x + y) \uparrow 2) + ((x - 4)/3)$.

(see Exercises 26 and 27). In particular, both a preorder traversal and a postorder traversal encode the structure of a full ordered m-ary tree. However, when the number of children of vertices is not specified, neither a preorder traversal nor a postorder traversal encodes the structure of an ordered rooted tree (see Exercises 28 and 29).

Infix, Prefix, and Postfix Notation

We can represent complicated expressions, such as compound propositions, combinations of sets, and arithmetic expressions using ordered rooted trees. For instance, consider the representation of an arithmetic expression involving the operators $+$ (addition), $-$ (subtraction), $*$ (multiplication), / (division), and $\uparrow$ (exponentiation). We will use parentheses to indicate the order of the operations. An ordered rooted tree can be used to represent such expressions, where the internal vertices represent operations, and the leaves represent the variables or numbers. Each operation operates on its left and right subtrees (in that order).

EXAMPLE 5 What is the ordered rooted tree that represents the expression $((x + y) \uparrow 2) + ((x - 4)/3)$?

Solution: The binary tree for this expression can be built from the bottom up. First, a subtree for the expression $x + y$ is constructed. Then this is incorporated as part of the larger subtree representing $(x + y) \uparrow 2$. Also, a subtree for $x - 4$ is constructed, and then this is incorporated into a subtree representing $(x - 4)/3$. Finally the subtrees representing $(x + y) \uparrow 2$ and $(x - 4)/3$ are combined to form the ordered rooted tree representing $((x + y) \uparrow 2) + ((x - 4)/3)$. These steps are shown in Figure 10. ◀

An inorder traversal of the binary tree representing an expression produces the original expression with the elements and operations in the same order as they originally occurred, except for unary operations, which instead immediately follow their operands. For instance, inorder traversals of the binary trees in Figure 11, which represent the expressions $(x + y)/(x + 3)$, $(x + (y/x)) + 3$, and $x + (y/(x + 3))$, all lead to the infix expression $x + y/x + 3$. To make such expressions unambiguous it is necessary to include parentheses in the inorder traversal whenever we encounter an operation. The fully parenthesized expression obtained in this way is said to be in **infix form.**

We obtain the **prefix form** of an expression when we traverse its rooted tree in preorder. Expressions written in prefix form are said to be in **Polish notation,** which is named after the Polish logician Jan Łukasiewicz. An expression in prefix notation (where each operation has

FIGURE 11 Rooted Trees Representing $(x + y)/(x + 3)$, $(x + (y/x)) + 3$, and $x + (y/(x + 3))$.

a specified number of operands), is unambiguous, so no parentheses are needed in such an expression. The verification of this is left as an exercise for the reader.

EXAMPLE 6 What is the prefix form for $((x + y) \uparrow 2) + ((x - 4)/3)$?

Solution: We obtain the prefix form for this expression by traversing the binary tree that represents it, shown in Figure 10. This produces $+ \uparrow + x\ y\ 2\ /\ -\ x\ 4\ 3$. ◀

In the prefix form of an expression, a binary operator, such as $+$, precedes its two operands. Hence, we can evaluate an expression in prefix form by working from right to left. When we encounter an operator, we perform the corresponding operation with the two operands immediately to the right of this operand. Also, whenever an operation is performed, we consider the result a new operand.

EXAMPLE 7 What is the value of the prefix expression $+ - * 2\ 3\ 5 / \uparrow 2\ 3\ 4$?

Solution: The steps used to evaluate this expression by working right to left, and performing operations using the operands on the right, are shown in Figure 12. The value of this expression is 3. ◀

Links

We obtain the **postfix form** of an expression by traversing its binary tree in postorder. Expressions written in postfix form are said to be in **reverse Polish notation.** Expressions in reverse Polish notation are unambiguous, so parentheses are not needed. The verification of this is left to the reader.

EXAMPLE 8 What is the postfix form of the expression $((x + y) \uparrow 2) + ((x - 4)/3)$?

Solution: The postfix form of the expression is obtained by carrying out a postorder traversal of the binary tree for this expression, shown in Figure 10. This produces the postfix expression: $x\ y + 2 \uparrow x\ 4 - 3\ /\ +$. ◀

In the postfix form of an expression, a binary operator follows its two operands. So, to evaluate an expression from its postfix form, work from left to right, carrying out operations whenever an operator follows two operands. After an operation is carried out, the result of this operation becomes a new operand.

$$+ \quad - \quad * \quad 2 \quad 3 \quad 5 \quad / \quad \underbrace{\uparrow \quad 2 \quad 3}_{2 \uparrow 3 = 8} \quad 4$$

$$+ \quad - \quad * \quad 2 \quad 3 \quad 5 \quad \underbrace{/ \quad 8 \quad 4}_{8/4 = 2}$$

$$+ \quad - \quad \underbrace{* \quad 2 \quad 3}_{2 \quad 3 = 6} \quad 5 \quad 2$$

$$+ \quad \underbrace{- \quad 6 \quad 5}_{6-5=1} \quad 2$$

$$\underbrace{+ \quad 1 \quad 2}_{1+2=3}$$

Value of expression 3

FIGURE 12 Evaluating a Prefix Expression.

$$7 \quad \underbrace{2 \quad 3 \quad *}_{2*3=6} \quad - \quad 4 \quad \uparrow \quad 9 \quad 3 \quad / \quad +$$

$$\underbrace{7 \quad 6 \quad -}_{7-6=1} \quad 4 \quad \uparrow \quad 9 \quad 3 \quad / \quad +$$

$$\underbrace{1 \quad 4 \quad \uparrow}_{1^4=1} \quad 9 \quad 3 \quad / \quad +$$

$$1 \quad \underbrace{9 \quad 3 \quad /}_{9/3=3} \quad +$$

$$\underbrace{1 \quad 3 \quad +}_{1+3=4}$$

Value of expression: 4

FIGURE 13 Evaluating a Postfix Expression.

EXAMPLE 9 What is the value of the postfix expression $7\ 2\ 3 * - 4 \uparrow 9\ 3/+$?

Solution: The steps used to evaluate this expression by starting at the left and carrying out operations when two operands are followed by an operator are shown in Figure 13. The value of this expression is 4. ◀

Rooted trees can be used to represent other types of expressions, such as those representing compound propositions and combinations of sets. In these examples unary operators, such as the negation of a proposition, occur. To represent such operators and their operands, a vertex representing the operator and a child of this vertex representing the operand are used.

EXAMPLE 10 Find the ordered rooted tree representing the compound proposition $(\neg(p \wedge q)) \leftrightarrow (\neg p \vee \neg q)$. Then use this rooted tree to find the prefix, postfix, and infix forms of this expression.

Links

JAN ŁUKASIEWICZ (1878–1956) Jan Łukasiewicz was born into a Polish-speaking family in Lvov. At that time Lvov was part of Austria, but it is now in the Ukraine. His father was a captain in the Austrian army. Łukasiewicz became interested in mathematics while in high school. He studied mathematics and philosophy at the University of Lvov at both the undergraduate and graduate levels. After completing his doctoral work he became a lecturer there, and in 1911 he was appointed to a professorship. When the University of Warsaw was reopened as a Polish university in 1915, Łukasiewicz accepted an invitation to join the faculty. In 1919 he served as the Polish Minister of Education. He returned to the position of professor at Warsaw University where he remained from 1920 to 1939, serving as rector of the university twice.

Łukasiewicz was one of the cofounders of the famous Warsaw School of Logic. He published his famous text, *Elements of Mathematical Logic,* in 1928. With his influence, mathematical logic was made a required course for mathematics and science undergraduates in Poland. His lectures were considered excellent, even attracting students of the humanities.

Łukasiewicz and his wife experienced great suffering during World War II, which he documented in a posthumously published autobiography. After the war they lived in exile in Belgium. Fortunately, in 1949 he was offered a position at the Royal Irish Academy in Dublin.

Łukasiewicz worked on mathematical logic throughout his career. His work on a three-valued logic was an important contribution to the subject. Nevertheless, he is best known in the mathematical and computer science communities for his introduction of parenthesis-free notation, now called Polish notation.

FIGURE 14 Constructing the Rooted Tree for a Compound Proposition.

Extra Examples

Solution: The rooted tree for this compound proposition is constructed from the bottom up. First, subtrees for $\neg p$ and $\neg q$ are formed (where $\neg$ is considered a unary operator). Also, a subtree for $p \wedge q$ is formed. Then subtrees for $\neg(p \wedge q)$ and $(\neg p) \vee (\neg q)$ are constructed. Finally, these two subtrees are used to form the final rooted tree. The steps of this procedure are shown in Figure 14.

The prefix, postfix, and infix forms of this expression are found by traversing this rooted tree in preorder, postorder, and inorder (including parentheses), respectively. These traversals give $\leftrightarrow \neg \wedge pq \vee \neg p \neg q$, $pq \wedge \neg p \neg q \neg \vee \leftrightarrow$, and $(\neg(p \wedge q)) \leftrightarrow ((\neg p) \vee (\neg q))$, respectively. ◀

Because prefix and postfix expressions are unambiguous and because they can be evaluated easily without scanning back and forth, they are used extensively in computer science. Such expressions are especially useful in the construction of compilers.

Exercises

In Exercises 1–3 construct the universal address system for the given ordered rooted tree. Then use this to order its vertices using the lexicographic order of their labels.

1.

2.

3.

4. Suppose that the address of the vertex v in the ordered rooted tree T is 3.4.5.2.4.

a) At what level is v?
b) What is the address of the parent of v?
c) What is the least number of siblings v can have?
d) What is the smallest possible number of vertices in T if v has this address?
e) Find the other addresses that must occur.

5. Suppose that the vertex with the largest address in an

ordered rooted tree T has address 2.3.4.3.1. Is it possible to determine the number of vertices in T?

6. Can the leaves of an ordered rooted tree have the following list of universal addresses? If so, construct such an ordered rooted tree.

a) 1.1.1, 1.1.2, 1.2, 2.1.1.1, 2.1.2, 2.1.3, 2.2, 3.1.1, 3.1.2.1, 3.1.2.2, 3.2

b) 1.1, 1.2.1, 1.2.2, 1.2.3, 2.1, 2.2.1, 2.3.1, 2.3.2, 2.4.2.1, 2.4.2.2, 3.1, 3.2.1, 3.2.2

c) 1.1, 1.2.1, 1.2.2, 1.2.2.1, 1.3, 1.4, 2, 3.1, 3.2, 4.1.1.1

In Exercises 7–9 determine the order in which a preorder traversal visits the vertices of the given ordered rooted tree.

7.

8.

9.

10. In which order are the vertices of the ordered rooted tree in Exercise 7 visited using an inorder traversal?

11. In which order are the vertices of the ordered rooted tree in Exercise 8 visited using an inorder traversal?

12. In which order are the vertices of the ordered rooted tree in Exercise 9 visited using an inorder traversal?

13. In which order are the vertices of the ordered rooted tree in Exercise 7 visited using a postorder traversal?

14. In which order are the vertices of the ordered rooted tree in Exercise 8 visited using a postorder traversal?

15. In which order are the vertices of the ordered rooted tree in Exercise 9 visited using a postorder traversal?

16. **a)** Represent the expression $((x+2) \uparrow 3) * (y - (3+x)) - 5$ using a binary tree.
Write this expression in

b) prefix notation.

c) postfix notation.

d) infix notation.

17. **a)** Represent the expressions $(x + xy) + (x/y)$ and $x + ((xy + x)/y)$ using binary trees.
Write these expressions in

b) prefix notation.

c) postfix notation.

d) infix notation.

18. **a)** Represent the compound propositions $\neg(p \wedge q) \leftrightarrow (\neg p \vee \neg q)$ and $(\neg p \wedge (q \leftrightarrow \neg p)) \vee \neg q$ using ordered rooted trees.
Write these expressions in

b) prefix notation.

c) postfix notation.

d) infix notation.

19. **a)** Represent $(A \cap B) - (A \cup (B - A))$ using an ordered rooted tree.
Write this expression in

b) prefix notation.

c) postfix notation.

d) infix notation.

***20.** In how many ways can the string $\neg p \wedge q \leftrightarrow \neg p \vee \neg q$ be fully parenthesized to yield an infix expression?

***21.** In how many ways can the string $A \cap B - A \cap B - A$ be fully parenthesized to yield an infix expression?

22. Draw the ordered rooted tree corresponding to each of these arithmetic expressions written in prefix notation. Then write each expression using infix notation.

a) $+ * + - 5\ 3\ 2\ 1\ 4$

b) $\uparrow + 2\ 3 - 5\ 1$

c) $* / 9\ 3 + * 2\ 4 - 7\ 6$

23. What is the value of each of these prefix expressions?

a) $- * 2 / 8\ 4\ 3$

b) $\uparrow - * 3\ 3 * 4\ 2\ 5$

c) $+ - \uparrow 3\ 2 \uparrow 2\ 3 / 6 - 4\ 2$

d) $* + 3 + 3 \uparrow 3 + 3\ 3\ 3$

24. What is the value of each of these postfix expressions?

a) $5\ 2\ 1 - - 3\ 1\ 4 + + *$

b) $9\ 3 / 5 + 7\ 2 - *$

c) $3\ 2 * 2 \uparrow 5\ 3 - 8\ 4 / * -$

25. Construct the ordered rooted tree whose preorder traversal is $a, b, f, c, g, h, i, d, e, j, k, l$, where a has four children,

c has three children, j has two children, b and e have one child each, and all other vertices are leaves.

*26. Show that an ordered rooted tree is uniquely determined when a list of vertices generated by a preorder traversal of the tree and the number of children of each vertex are specified.

*27. Show that an ordered rooted tree is uniquely determined when a list of vertices generated by a postorder traversal of the tree and the number of children of each vertex are specified.

28. Show that preorder traversals of the two ordered rooted trees displayed below produce the same list of vertices. Note that this does not contradict the statement in Exercise 26, because the numbers of children of internal vertices in the two ordered rooted trees differ.

29. Show that postorder traversals of these two ordered rooted trees produce the same list of vertices. Note that this does not contradict the statement in Exercise 27, because the numbers of children of internal vertices in the two ordered rooted trees differ.

Well-formed formulae in prefix notation over a set of symbols and a set of binary operators are defined recursively by these rules:

(*i*) if x is a symbol, then x is a well-formed formula in prefix notation;

(*ii*) if X and Y are well-formed formulae and $*$ is an operator, then $* XY$ is a well-formed formula.

30. Which of these are well-formed formulae over the symbols $\{x, y, z\}$ and the set of binary operators $\{\times, +, \circ\}$?

a) $\times + + x\, y\, x$
b) $\circ\, x\, y \times x\, z$
c) $\times \circ x\, z \times \times x\, y$
d) $\times + \circ x\, x \circ x\, x\, x$

*31. Show that any well-formed formula in prefix notation over a set of symbols and a set of binary operators contains exactly one more symbol than the number of operators.

32. Give a definition of well-formed formulae in postfix notation over a set of symbols and a set of binary operators.

33. Give six examples of well-formed formulae with three or more operators in postfix notation over the set of symbols $\{x, y, z\}$ and the set of operators $\{+, \times, \circ\}$.

34. Extend the definition of well-formed formulae in prefix notation to sets of symbols and operators where the operators may not be binary.

10.4 Spanning Trees

Introduction

Consider the system of roads in Maine represented by the simple graph shown in Figure 1(a). The only way the roads can be kept open in the winter is by frequently plowing them. The highway department wants to plow the fewest roads so that there will always be cleared roads connecting any two towns. How can this be done?

At least five roads must be plowed to ensure that there is a path between any two towns. Figure 1(b) shows one such set of roads. Note that the subgraph representing these roads is a tree, because it is connected and contains six vertices and five edges.

This problem was solved with a connected subgraph with the minimum number of edges containing all vertices of the original simple graph. Such a graph must be a tree.

DEFINITION 1 Let G be a simple graph. A *spanning tree* of G is a subgraph of G that is a tree containing every vertex of G.

FIGURE 1 (a) A Road System and (b) a Set of Roads to Plow.

FIGURE 2 The Simple Graph G.

A simple graph with a spanning tree must be connected, because there is a path in the spanning tree between any two vertices. The converse is also true; that is, every connected simple graph has a spanning tree. We will give an example before proving this result.

EXAMPLE 1 Find a spanning tree of the simple graph G shown in Figure 2.

Solution: The graph G is connected, but it is not a tree because it contains simple circuits. Remove the edge $\{a, e\}$. This eliminates one simple circuit, and the resulting subgraph is still connected and still contains every vertex of G. Next remove the edge $\{e, f\}$ to eliminate a second simple circuit. Finally, remove edge $\{c, g\}$ to produce a simple graph with no simple circuits. This subgraph is a spanning tree, because it is a tree that contains every vertex of G. The sequence of edge removals used to produce the spanning tree is illustrated in Figure 3.

The tree shown in Figure 3 is not the only spanning tree of G. For instance, each of the trees shown in Figure 4 is a spanning tree of G. ◀

THEOREM 1 A simple graph is connected if and only if it has a spanning tree.

Proof: First, suppose that a simple graph G has a spanning tree T. T contains every vertex of G. Furthermore, there is a path in T between any two of its vertices. Because T is a subgraph of G, there is a path in G between any two of its vertices. Hence, G is connected.

Now suppose that G is connected. If G is not a tree, it must contain a simple circuit. Remove an edge from one of these simple circuits. The resulting subgraph has one fewer edge but still contains all the vertices of G and is connected. This subgraph is still connected because when two vertices are connected by a path containing the removed edge, they are connected by a path not containing this edge. We can construct such a path by inserting into the original path, at

FIGURE 3 Producing a Spanning Tree for G by Removing Edges That Form Simple Circuits.

FIGURE 4 Spanning Trees of *G*.

the point where the removed edge once was, the simple circuit with this edge removed. If this subgraph is not a tree, it has a simple circuit; so as before, remove an edge that is in a simple circuit. Repeat this process until no simple circuits remain. This is possible because there are only a finite number of edges in the graph. The process terminates when no simple circuits remain. A tree is produced because the graph stays connected as edges are removed. This tree is a spanning tree because it contains every vertex of G. ◁

Spanning trees are important in data networking, as Example 2 shows.

EXAMPLE 2 **IP Multicasting** Spanning trees play an important role in multicasting over Internet Protocol (IP) networks. To send data from a source computer to multiple receiving computers, each of which is a subnetwork, data could be sent separately to each computer. This type of networking, called unicasting, is inefficient, because many copies of the same data are transmitted over the network. To make the transmission of data to multiple receiving computers more efficient, IP multicasting is used. With IP multicasting, a computer sends a single copy of data over the network, and as data reaches intermediate routers, the data are forwarded to one or more other routers so that ultimately all receiving computers in their various subnetworks receive these data. (Routers are computers that are dedicated to forwarding IP datagrams between subnetworks in a network. In multicasting, routers use Class D addresses, each representing a session that receiving computers may join; see Example 16 in Section 5.1.)

For data to reach receiving computers as quickly as possible, there should be no loops (which in graph theory terminology are circuits or cycles) in the path that data take through the network. That is, once data have reached a particular router, data should never return to this router. To avoid loops, the multicast routers use network algorithms to construct a spanning tree in the graph that has the multicast source, the routers, and the subnetworks containing receiving computers as vertices, with edges representing the links between computers and/or routers. The root of this spanning tree is the multicast source. The subnetworks containing receiving computers are leaves of the tree. (Note that subnetworks not containing receiving stations are not included in the graph.) This is illustrated in Figure 5. ◀

Depth-First Search

The proof of Theorem 1 gives an algorithm for finding spanning trees by removing edges from simple circuits. This algorithm is inefficient, because it requires that simple circuits be identified.

FIGURE 5 A Multicast Spanning Tree.

Instead of constructing spanning trees by removing edges, spanning trees can be built up by successively adding edges. Two algorithms based on this principle will be presented here.

Demo

We can build a spanning tree for a connected simple graph using **depth-first search.** We will form a rooted tree, and the spanning tree will be the underlying undirected graph of this rooted tree. Arbitrarily choose a vertex of the graph as the root. Form a path starting at this vertex by successively adding vertices and edges, where each new edge is incident with the last vertex in the path and a vertex not already in the path. Continue adding vertices and edges to this path as long as possible. If the path goes through all vertices of the graph, the tree consisting of this path is a spanning tree. However, if the path does not go through all vertices, more vertices and edges must be added. Move back to the next to last vertex in the path, and, if possible, form a new path starting at this vertex passing through vertices that were not already visited. If this cannot be done, move back another vertex in the path, that is, two vertices back in the path, and try again.

Repeat this procedure, beginning at the last vertex visited, moving back up the path one vertex at a time, forming new paths that are as long as possible until no more edges can be added. Because the graph has a finite number of edges and is connected, this process ends with the production of a spanning tree. Each vertex that ends a path at a stage of the algorithm will be a leaf in the rooted tree, and each vertex where a path is constructed starting at this vertex will be an internal vertex.

The reader should note the recursive nature of this procedure. Also, note that if the vertices in the graph are ordered, the choices of edges at each stage of the procedure are all determined when we always choose the first vertex in the ordering that is available. However, we will not always explicitly order the vertices of a graph.

Depth-first search is also called **backtracking,** because the algorithm returns to vertices previously visited to add paths. Example 3 illustrates backtracking.

EXAMPLE 3 Use depth-first search to find a spanning tree for the graph G shown in Figure 6.

Extra Examples

Solution: The steps used by depth-first search to produce a spanning tree of G are shown in Figure 7. We arbitrarily start with the vertex f. A path is built by successively adding edges incident with vertices not already in the path, as long as this is possible. This produces a path f, g, h, k, j (note that other paths could have been built). Next, backtrack to k. There is no path

FIGURE 6 The Graph G.

FIGURE 7 Depth-First Search of G.

beginning at k containing vertices not already visited. So we backtrack to h. Form the path h, i. Then backtrack to h, and then to f. From f build the path f, d, e, c, a. Then backtrack to c and form the path c, b. This produces the spanning tree. ◀

The edges selected by depth-first search of a graph are called **tree edges.** All other edges of the graph must connect a vertex to an ancestor or descendant of this vertex in the tree. These edges are called **back edges.** (Exercise 39 asks for a proof of this fact.)

EXAMPLE 4 In Figure 8 we highlight the tree edges found by depth-first search starting at vertex f by showing them with heavy colored lines. The back edges (e, f) and (f, h) are shown with thinner black lines. ◀

We have explained how to find a spanning tree of a graph using depth-first search. However, our discussion so far has not brought out the recursive nature of depth-first search. To help make the recursive nature of the algorithm clear, we need a little terminology. We say that we *explore* from a vertex v when we carry out the steps of depth-first search beginning when v is added to the tree and ending when we have backtracked back to v for the last time. The key observation needed to understand the recursive nature of the algorithm is that when we add an edge connecting a vertex v to a vertex w, we finish exploring from w before we return to v to complete exploring from v.

In Algorithm 1 we construct the spanning tree of a graph G with vertices $v_1, \ldots, v_n$ by first selecting the vertex v_1 to be the root. We initially set T to be the tree with just this one vertex. At each step we add a new vertex to the tree T together with an edge from a vertex already in T to this new vertex and we explore from this new vertex. Note that at the completion of the algorithm, T contains no simple circuits because no edge is added that connects a vertex already in the tree. Moreover, T remains connected as it is built. (These last two observations can be easily proved via mathematical induction.) Because G is connected, every vertex in G is visited by the algorithm and is added to the tree (as the reader should verify). It follows that T is a spanning tree of G.

FIGURE 8 The Tree Edges and Back Edges of the Depth-First Search in Example 4.

ALGORITHM 1 **Depth-First Search.**

procedure *DFS*(G: connected graph with vertices $v_1, v_2, \ldots, v_n$)
$T :=$ tree consisting only of the vertex v_1
visit(v_1)
procedure *visit*(v: vertex of G)
for each vertex w adjacent to v and not yet in T
begin
  add vertex w and edge $\{v, w\}$ to T
  visit(w)
end

We now analyze the computational complexity of the depth-first search algorithm. The key observation is that for each vertex v, the procedure *visit*(v) is called when the vertex v is first encountered in the search and it is not called again. Assuming that the adjacency lists for G are available (see Section 9.3), no computations are required to find the vertices adjacent to v. As we follow the steps of the algorithm, we examine each edge at most twice to determine whether to add this edge and one of its endpoints to the tree. Consequently, the procedure *DFS* constructs a spanning tree using $O(e)$, or $O(n^2)$, steps where e and n are the number of edges and vertices in G, respectively. [Note that a step involves examining a vertex to see whether it is already in the spanning tree as it is being built and adding this vertex and the corresponding edge if the vertex is not already in the tree. We have also made use of the inequality $e \leq n(n-1)/2$, which holds for any simple graph.]

Depth-first search can be used as the basis for algorithms that solve many different problems. For example, it can be used to find paths and circuits in a graph, it can be used to determine the connected components of a graph, and it can be used to find the cut vertices of a connected graph. As we will see, depth-first search is the basis of backtracking techniques used to search for solutions of computationally difficult problems. (See [GrYe99], [Ma89], and [CoLeRiSt01] for a discussion of algorithms based on depth-first search.)

Breadth-First Search

Demo Links

We can also produce a spanning tree of a simple graph by the use of **breadth-first search.** Again, a rooted tree will be constructed, and the underlying undirected graph of this rooted tree forms the spanning tree. Arbitrarily choose a root from the vertices of the graph. Then add all edges incident to this vertex. The new vertices added at this stage become the vertices at level 1 in the spanning tree. Arbitrarily order them. Next, for each vertex at level 1, visited in order, add each edge incident to this vertex to the tree as long as it does not produce a simple circuit. Arbitrarily order the children of each vertex at level 1. This produces the vertices at level 2 in the tree. Follow the same procedure until all the vertices in the tree have been added. The procedure ends because there are only a finite number of edges in the graph. A spanning tree is produced because we have produced a tree containing every vertex of the graph. An example of breadth-first search is given in Example 5.

EXAMPLE 5 Use breadth-first search to find a spanning tree for the graph shown in Figure 9.

Extra Examples

Solution: The steps of the breadth-first search procedure are shown in Figure 10. We choose the vertex e to be the root. Then we add edges incident with all vertices adjacent to e, so edges from e to b, d, f, and i are added. These four vertices are at level 1 in the tree. Next, add the edges from these vertices at level 1 to adjacent vertices not already in the tree. Hence, the edges

FIGURE 9 A Graph *G*.

from b to a and c are added, as are edges from d to h, from f to j and g, and from i to k. The new vertices a, c, h, j, g, and k are at level 2. Next, add edges from these vertices to adjacent vertices not already in the graph. This adds edges from g to l and from k to m. ◀

We describe breadth-first search in pseudocode as Algorithm 2. In this algorithm, we assume the vertices of the connected graph G are ordered as $v_1, v_2, \ldots, v_n$. In the algorithm we use the term "process" to describe the procedure of adding new vertices, and corresponding edges, to the tree adjacent to the current vertex being processed as long as a simple circuit is not produced.

ALGORITHM 2 **Breadth-First Search.**

```
procedure BFS (G: connected graph with vertices v1, v2, ..., vn)
T := tree consisting only of vertex v1
L := empty list
put v1 in the list L of unprocessed vertices
while L is not empty
begin
   remove the first vertex, v, from L
   for each neighbor w of v
      if w is not in L and not in T then
      begin
         add w to the end of the list L
         add w and edge {v, w} to T
      end
end
```

FIGURE 10 Breadth-First Search of *G*.

We now analyze the computational complexity of breadth-first search. For each vertex v in the graph we examine all vertices adjacent to v and we add each vertex not yet visited to the tree T. Assuming we have the adjacency lists for the graph available, no computation is required to determine which vertices are adjacent to a given vertex. As in the analysis of the depth-first search algorithm, we see that we examine each edge at most twice to determine whether we should add this edge and its endpoint not already in the tree. It follows that the breadth-first search algorithm uses $O(e)$ or $O(n^2)$ steps.

Backtracking Applications

There are problems that can be solved only by performing an exhaustive search of all possible solutions. One way to search systematically for a solution is to use a decision tree, where each internal vertex represents a decision and each leaf a possible solution. To find a solution via backtracking, first make a sequence of decisions in an attempt to reach a solution as long as this is possible. The sequence of decisions can be represented by a path in the decision tree. Once it is known that no solution can result from any further sequence of decisions, backtrack to the parent of the current vertex and work toward a solution with another series of decisions, if this is possible. The procedure continues until a solution is found, or it is established that no solution exists. Examples 6 to 8 illustrate the usefulness of backtracking.

EXAMPLE 6 **Graph Colorings** How can backtracking be used to decide whether a graph can be colored using n colors?

Solution: We can solve this problem using backtracking in the following way. First pick some vertex a and assign it color 1. Then pick a second vertex b, and if b is not adjacent to a, assign it color 1. Otherwise, assign color 2 to b. Then go on to a third vertex c. Use color 1, if possible, for c. Otherwise use color 2, if this is possible. Only if neither color 1 nor color 2 can be used should color 3 be used. Continue this process as long as it is possible to assign one of the n colors to each additional vertex, always using the first allowable color in the list. If a vertex is reached that cannot be colored by any of the n colors, backtrack to the last assignment made and change the coloring of the last vertex colored, if possible, using the next allowable color in the list. If it is not possible to change this coloring, backtrack farther to previous assignments, one step back at a time, until it is possible to change a coloring of a vertex. Then continue assigning colors of additional vertices as long as possible. If a coloring using n colors exists, backtracking will produce it. (Unfortunately this procedure can be extremely inefficient.)

In particular, consider the problem of coloring the graph shown in Figure 11 with three colors. The tree shown in Figure 11 illustrates how backtracking can be used to construct a 3-coloring. In this procedure, red is used first, then blue, and finally green. This simple example can obviously be done without backtracking, but it is a good illustration of the technique.

In this tree, the initial path from the root, which represents the assignment of red to a, leads to a coloring with a red, b blue, c red, and d green. It is impossible to color e using any of the three colors when a, b, c, and d are colored in this way. So, backtrack to the parent of the vertex representing this coloring. Because no other color can be used for d, backtrack one more level. Then change the color of c to green. We obtain a coloring of the graph by then assigning red to d and green to e. ◀

EXAMPLE 7 **The n-Queens Problem** The n-queens problem asks how n queens can be placed on an $n \times n$ chessboard so that no two queens can attack one another. How can backtracking be used to solve the n-queens problem?

Links

Solution: To solve this problem we must find n positions on an $n \times n$ chessboard so that no two of these positions are in the same row, same column, or in the same diagonal [a diagonal consists of all positions (i, j) with $i + j = m$ for some m, or $i - j = m$ for some m]. We will

FIGURE 11 Coloring a Graph Using Backtracking.

use backtracking to solve the n-queens problem. We start with an empty chessboard. At stage $k+1$ we attempt putting an additional queen on the board in the $(k+1)$st column, where there are already queens in the first k columns. We examine squares in the $(k+1)$st column starting with the square in the first row, looking for a position to place this queen so that it is not in the same row or on the same diagonal as a queen already on the board. (We already know it is not in the same column.) If it is impossible to find a position to place the queen in the $(k+1)$st column, backtrack to the placement of the queen in the kth column, and place this queen in the next allowable row in this column, if such a row exists. If no such row exists, backtrack further.

In particular, Figure 12 displays a backtracking solution to the four-queens problem. In this solution, we place a queen in the first row and column. Then we put a queen in the third row of the second column. However, this makes it impossible to place a queen in the third column. So we backtrack and put a queen in the fourth row of the second column. When we do this, we can place a queen in the second row of the third column. But there is no way to add a queen to the fourth column. This shows that no solution results when a queen is placed in the first row and column. We backtrack to the empty chessboard, and place a queen in the second row of the first column. This leads to a solution as shown in Figure 12. ◀

EXAMPLE 8 **Sums of Subsets** Consider this problem. Given a set of positive integers $x_1, x_2, \ldots, x_n$, find a subset of this set of integers that has M as its sum. How can backtracking be used to solve this problem?

Solution: We start with a sum with no terms. We build up the sum by successively adding terms. An integer in the sequence is included if the sum remains less than M when this integer is added to the sum. If a sum is reached such that the addition of any term is greater than M, backtrack by dropping the last term of the sum.

Figure 13 displays a backtracking solution to the problem of finding a subset of $\{31, 27, 15, 11, 7, 5\}$ with the sum equal to 39. ◀

Depth-First Search in Directed Graphs

We can easily modify both depth-first search and breadth-first search so that they can run given a directed graph as input. However, the output will not necessarily be a spanning tree, but rather a spanning forest. In both algorithms we can add an edge only when it is directed away from

FIGURE 12 A Backtracking Solution of the Four-Queens Problem.

the vertex that is being visited and to a vertex not yet added. If at a stage of either algorithm we find that no edge exists starting at a vertex already added to one not yet added, the next vertex added by the algorithm becomes the root of a new tree in the spanning forest. This is illustrated in Example 9.

EXAMPLE 9 What is the output of depth-first search given the graph G shown in Figure 14(a) as input?

Solution: We begin the depth-first search at vertex a and add vertices b, c, and g and the corresponding edges where we are blocked. We backtrack to c but we are still blocked, and then

FIGURE 13 Find a Sum Equal to 39 Using Backtracking.

FIGURE 14 Depth-First Search of a Directed Graph.

backtrack to b, where we add vertices f and e and the corresponding edges. Backtracking takes us all the way back to a. We then start a new tree at d and add vertices h, l, k, and j and the corresponding edges. We backtrack to k, then l, then h, and back to d. Finally, we start a new tree at i, completing the depth-first search. The output is shown in Figure 14(b). ◀

Depth-first search in directed graphs is the basis of many algorithms (see [GrYe99], [Ma89], and [CoLeRiSt01]). It can be used to determine whether a directed graph has a circuit, it can be used to carry out a topological sort of a graph, and it can also be used to find the strongly connected components of a directed graph.

We conclude this section with an application of depth-first search and breadth-first search to search engines on the Web.

EXAMPLE 10 **Web Spiders** To index websites, search engines such as Google and Yahoo systematically explore the Web starting at known sites. These search engines use programs called Web spiders (or crawlers or bots) to visit websites and analyze their contents. Web spiders use both depth-first searching and breadth-first searching to create indices. As described in Example 8 in Section 9.1, Web pages and links between them can be modeled by a directed graph called the Web graph. Web pages are represented by vertices and links are represented by directed edges. Using depth-first search, an initial Web page is selected, a link is followed to a second Web page (if there is such a link), a link on the second Web page is followed to a third Web page, if there is such a link, and so on, until a page with no new links is found. Backtracking is then used to examine links at the previous level to look for new links, and so on. (Because of practical limitations, Web spiders have limits to the depth they search in depth-first search.) Using breadth-first search, an initial Web page is selected and a link on this page is followed to a second Web page, then a second link on the initial page is followed (if it exists), and so on, until all links of the initial page have been followed. Then links on the pages one level down are followed, page by page, and so on. ◀

Exercises

1. How many edges must be removed from a connected graph with n vertices and m edges to produce a spanning tree?

In Exercises 2–6 find a spanning tree for the graph shown by removing edges in simple circuits.

2.

3.

4.

5.

6.

7. Find a spanning tree for each of these graphs.

a) K_5 **b)** $K_{4,4}$ **c)** $K_{1,6}$
d) Q_3 **e)** C_5 **f)** W_5

In Exercises 8–10 draw all the spanning trees of the given simple graphs.

8.

9.

10.

***11.** How many different spanning trees does each of these simple graphs have?

a) K_3 **b)** K_4 **c)** $K_{2,2}$ **d)** C_5

***12.** How many nonisomorphic spanning trees does each of these simple graphs have?

a) K_3 **b)** K_4 **c)** K_5

In Exercises 13–15 use depth-first search to produce a spanning tree for the given simple graph. Choose a as the root of this spanning tree and assume that the vertices are ordered alphabetically.

13.

14.

15.

16. Use breadth-first search to produce a spanning tree for each of the simple graphs in Exercises 13–15. Choose a as the root of each spanning tree.

17. Use depth-first search to find a spanning tree of each of these graphs.

a) W_6 (see Example 7 of Section 9.2), starting at the vertex of degree 6
b) K_5
c) $K_{3,4}$, starting at a vertex of degree 3
d) Q_3

18. Use breadth-first search to find a spanning tree of each of the graphs in Exercise 17.

19. Describe the trees produced by breadth-first search and depth-first search of the wheel graph W_n, starting at the

vertex of degree n, where n is an integer with $n \geq 3$. (See Example 7 of Section 9.2.) Justify your answers.

20. Describe the trees produced by breadth-first search and depth-first search of the complete graph K_n, where n is a positive integer. Justify your answers.

21. Describe the trees produced by breadth-first search and depth-first search of the complete bipartite graph $K_{m,n}$, starting at a vertex of degree m, where m and n are positive integers. Justify your answers.

22. Explain how breadth-first search and how depth-first search can be used to determine whether a graph is bipartite.

23. Suppose that an airline must reduce its flight schedule to save money. If its original routes are as illustrated here, which flights can be discontinued to retain service between all pairs of cities (where it may be necessary to combine flights to fly from one city to another)?

24. When must an edge of a connected simple graph be in every spanning tree for this graph?

25. Which connected simple graphs have exactly one spanning tree?

26. Explain how breadth-first search or depth-first search can be used to order the vertices of a connected graph.

***27.** Show that the length of the shortest path between vertices v and u in a connected simple graph equals the level number of u in the breadth-first spanning tree of G with root v.

28. Use backtracking to try to find a coloring of each of the graphs in Exercises 7–9 of Section 9.8 using three colors.

29. Use backtracking to solve the n-queens problem for these values of n.

a) $n = 3$ **b)** $n = 5$ **c)** $n = 6$

30. Use backtracking to find a subset, if it exists, of the set $\{27, 24, 19, 14, 11, 8\}$ with sum

a) 20. **b)** 41. **c)** 60.

31. Explain how backtracking can be used to find a Hamilton path or circuit in a graph.

32. a) Explain how backtracking can be used to find the way out of a maze, given a starting position and the exit position. Consider the maze divided into positions, where at each position the set of available moves includes one to four possibilities (up, down, right, left).

b) Find a path from the starting position marked by X to the exit in this maze.

A **spanning forest** of graph G is a forest that contains every vertex of G such that two vertices are in the same tree of the forest when there is a path in G between these two vertices.

33. Show that every finite simple graph has a spanning forest.

34. How many trees are in the spanning forest of a graph?

35. How many edges must be removed to produce the spanning forest of a graph with n vertices, m edges, and c connected components?

36. Devise an algorithm for constructing the spanning forest of a graph based on deleting edges that form simple circuits.

37. Devise an algorithm for constructing the spanning forest of a graph based on depth-first searching.

38. Devise an algorithm for constructing the spanning forest of a graph based on breadth-first searching.

39. Let G be a connected graph. Show that if T is a spanning tree of G constructed using depth-first search, then an edge of G not in T must be a back edge, that is, it must connect a vertex to one of its ancestors or one of its descendants in T.

40. Let G be a connected graph. Show that if T is a spanning tree of G constructed using breadth-first search, then an edge of G not in T must connect vertices at the same level or at levels that differ by 1 in this spanning tree.

41. For which graphs do depth-first search and breadth-first search produce identical spanning trees no matter which vertex is selected as the root of the tree? Justify your answer.

42. Use Exercise 39 to prove that if G is a connected, simple graph with n vertices and G does not contain a simple path of length k then it contains at most $(k-1)n$ edges.

43. Use mathematical induction to prove that breadth-first search visits vertices in order of their level in the resulting spanning tree.

44. Use pseudocode to describe a variation of depth-first search that assigns the integer n to the nth vertex visited in the search. Show that this numbering corresponds to the numbering of the vertices created by a preorder traversal of the spanning tree.

45. Use pseudocode to describe a variation of breadth-first search that assigns the integer m to the mth vertex visited in the search.

***46.** Suppose that G is a directed graph and T is a spanning tree constructed using breadth-first search. Show that every edge of G connects two vertices at the same level, a vertex to a vertex at one level lower, or a vertex to a vertex at some higher level.

47. Show that if G is a directed graph and T is a spanning tree constructed using depth-first search, then every edge not in the spanning tree is a **forward edge** connecting an ancestor to a descendant, a **back edge** connecting a descendant to an ancestor, or a **cross edge** connecting a vertex to a vertex in a previously visited subtree.

***48.** Describe a variation of depth-first search that assigns the smallest available positive integer to a vertex when the algorithm is totally finished with this vertex. Show that in this numbering, each vertex has a larger number than its children and that the children have increasing numbers from left to right.

Let T_1 and T_2 be spanning trees of a graph. The **distance** between T_1 and T_2 is the number of edges in T_1 and T_2 that are not common to T_1 and T_2.

49. Find the distance between each pair of spanning trees shown in Figures 3(c) and 4 of the graph G shown in Figure 2.

***50.** Suppose that T_1, T_2, and T_3 are spanning trees of the simple graph G. Show that the distance between T_1 and T_3 does not exceed the sum of the distance between T_1 and T_2 and the distance between T_2 and T_3.

****51.** Suppose that T_1 and T_2 are spanning trees of a simple graph G. Moreover, suppose that e_1 is an edge in T_1 that is not in T_2. Show that there is an edge e_2 in T_2 that is not in T_1 such that T_1 remains a spanning tree if e_1 is removed from it and e_2 is added to it, and T_2 remains a spanning tree if e_2 is removed from it and e_1 is added to it.

***52.** Show that it is possible to find a sequence of spanning trees leading from any spanning tree to any other by successively removing one edge and adding another.

A **rooted spanning tree** of a directed graph is a rooted tree containing edges of the graph such that every vertex of the graph is an endpoint of one of the edges in the tree.

53. For each of the directed graphs in Exercises 18–23 of Section 9.5 either find a rooted spanning tree of the graph or determine that no such tree exists.

***54.** Show that a connected directed graph in which each vertex has the same in-degree and out-degree has a rooted spanning tree. [*Hint:* Use an Euler circuit.]

***55.** Give an algorithm to build a rooted spanning tree for connected directed graphs in which each vertex has the same in-degree and out-degree.

***56.** Show that if G is a directed graph and T is a spanning tree constructed using depth-first search, then G contains a circuit if and only if G contains a back edge (see Exercise 47) relative to the spanning tree T.

***57.** Use Exercise 56 to construct an algorithm for determining whether a directed graph contains a circuit.

10.5 Minimum Spanning Trees

Introduction

Links

A company plans to build a communications network connecting its five computer centers. Any pair of these centers can be linked with a leased telephone line. Which links should be made to ensure that there is a path between any two computer centers so that the total cost of the network is minimized? We can model this problem using the weighted graph shown in Figure 1, where vertices represent computer centers, edges represent possible leased lines, and the weights on edges are the monthly lease rates of the lines represented by the edges. We can solve this problem by finding a spanning tree so that the sum of the weights of the edges of the tree is minimized. Such a spanning tree is called a **minimum spanning tree.**

FIGURE 1 A Weighted Graph Showing Monthly Lease Costs for Lines in a Computer Network.

Algorithms for Minimum Spanning Trees

A wide variety of problems are solved by finding a spanning tree in a weighted graph such that the sum of the weights of the edges in the tree is a minimum.

DEFINITION 1 A *minimum spanning tree* in a connected weighted graph is a spanning tree that has the smallest possible sum of weights of its edges.

Demo

We will present two algorithms for constructing minimum spanning trees. Both proceed by successively adding edges of smallest weight from those edges with a specified property that have not already been used. Both are greedy algorithms. Recall from Section 3.1 that a greedy algorithm is a procedure that makes an optimal choice at each of its steps. Optimizing at each step does not guarantee that the optimal overall solution is produced. However, the two algorithms presented in this section for constructing minimum spanning trees are greedy algorithms that do produce optimal solutions.

Links

The first algorithm that we will discuss was given by Robert Prim in 1957, although the basic ideas of this algorithm have an earlier origin. To carry out **Prim's algorithm,** begin by choosing any edge with smallest weight, putting it into the spanning tree. Successively add to the tree edges of minimum weight that are incident to a vertex already in the tree and not forming a simple circuit with those edges already in the tree. Stop when $n - 1$ edges have been added.

Later in this section, we will prove that this algorithm produces a minimum spanning tree for any connected weighted graph. Algorithm 1 gives a pseudocode description of Prim's algorithm.

ALGORITHM 1 **Prim's Algorithm.**

procedure *Prim*(G: weighted connected undirected graph with n vertices)
T := a minimum-weight edge
for i := 1 **to** $n - 2$
begin
  e := an edge of minimum weight incident to a vertex in T and not forming a simple circuit in T if added to T
  T := T with e added
end {T is a minimum spanning tree of G}

Note that the choice of an edge to add at a stage of the algorithm is not determined when there is more than one edge with the same weight that satisfies the appropriate criteria. We need to order the edges to make the choices deterministic. We will not worry about this in the remainder of the section. Also note that there may be more than one minimum spanning tree for a given

ROBERT CLAY PRIM (BORN 1921) Robert Prim, born in Sweetwater, Texas, received his B.S. in electrical engineering in 1941 and his Ph.D. in mathematics from Princeton University in 1949. He was an engineer at the General Electric Company from 1941 until 1944, an engineer and mathematician at the United States Naval Ordnance Lab from 1944 until 1949, and a research associate at Princeton University from 1948 until 1949. Among the other positions he has held are director of mathematics and mechanics research at Bell Telephone Laboratories from 1958 until 1961 and vice president of research at Sandia Corporation. He is currently retired.

Choice	Edge	Cost
1	{Chicago, Atlanta}	$ 700
2	{Atlanta, New York}	$ 800
3	{Chicago, San Francisco}	$ 1200
4	{San Francisco, Denver}	$ 900
	Total:	$ 3600

FIGURE 2 A Minimum Spanning Tree for the Weighted Graph in Figure 1.

FIGURE 3 A Weighted Graph.

connected weighted simple graph. (See Exercise 9.) Examples 1 and 2 illustrate how Prim's algorithm is used.

EXAMPLE 1 Use Prim's algorithm to design a minimum-cost communications network connecting all the computers represented by the graph in Figure 1.

Solution: We solve this problem by finding a minimum spanning tree in the graph in Figure 1. Prim's algorithm is carried out by choosing an initial edge of minimum weight and successively adding edges of minimum weight that are incident to a vertex in the tree and that do not form simple circuits. The edges in color in Figure 2 show a minimum spanning tree produced by Prim's algorithm, with the choice made at each step displayed. ◀

EXAMPLE 2 Use Prim's algorithm to find a minimum spanning tree in the graph shown in Figure 3.

Solution: A minimum spanning tree constructed using Prim's algorithm is shown in Figure 4. The successive edges chosen are displayed. ◀

Links

The second algorithm we will discuss was discovered by Joseph Kruskal in 1956, although the basic ideas it uses were described much earlier. To carry out **Kruskal's algorithm,** choose an edge in the graph with minimum weight.

Successively add edges with minimum weight that do not form a simple circuit with those edges already chosen. Stop after $n - 1$ edges have been selected.

The proof that Kruskal's algorithm produces a minimum spanning tree for every connected weighted graph is left as an exercise at the end of this section. Pseudocode for Kruskal's algorithm is given in Algorithm 2.

Links

JOSEPH BERNARD KRUSKAL (BORN 1928) Joseph Kruskal, born in New York City, attended the University of Chicago and received his Ph.D. from Princeton University in 1954. He was an instructor in mathematics at Princeton and at the University of Wisconsin, and later he was an assistant professor at the University of Michigan. In 1959 he became a member of the technical staff at Bell Laboratories, a position he continues to hold. His current research interests include statistical linguistics and psychometrics. Besides his work on minimum spanning trees, Kruskal is also known for contributions to multidimensional scaling. Kruskal discovered his algorithm for producing minimum spanning trees when he was a second-year graduate student. He was not sure his $2\frac{1}{2}$-page paper on this subject was worthy of publication, but was convinced by others to submit it.

(a)

Choice	Edge	Weight
1	$\{b, f\}$	1
2	$\{a, b\}$	2
3	$\{f, j\}$	2
4	$\{a, e\}$	3
5	$\{i, j\}$	3
6	$\{f, g\}$	3
7	$\{c, g\}$	2
8	$\{c, d\}$	1
9	$\{g, h\}$	3
10	$\{h, l\}$	3
11	$\{k, l\}$	1
	Total:	24

(b)

FIGURE 4 A Minimum Spanning Tree Produced Using Prim's Algorithm.

ALGORITHM 2 **Kruskal's Algorithm.**

procedure *Kruskal*(G: weighted connected undirected graph with n vertices)
T := empty graph
for i := 1 **to** $n - 1$
begin
  e := any edge in G with smallest weight that does not form a simple circuit when added to T
  T := T with e added
end {T is a minimum spanning tree of G}

The reader should note the difference between Prim's and Kruskal's algorithms. In Prim's algorithm edges of minimum weight that are incident to a vertex already in the tree, and not forming a circuit, are chosen; whereas in Kruskal's algorithm edges of minimum weight that are not necessarily incident to a vertex already in the tree, and that do not form a circuit, are chosen. Note that as in Prim's algorithm, if the edges are not ordered, there may be more than one choice for the edge to add at a stage of this procedure. Consequently, the edges need to be ordered for the procedure to be deterministic. Example 3 illustrates how Kruskal's algorithm is used.

EXAMPLE 3 Use Kruskal's algorithm to find a minimum spanning tree in the weighted graph shown in Figure 3.

Extra Examples

Solution: A minimum spanning tree and the choices of edges at each stage of Kruskal's algorithm are shown in Figure 5. ◀

HISTORICAL NOTE Joseph Kruskal and Robert Prim developed their algorithms for constructing minimum spanning trees in the mid-1950s. However, they were not the first people to discover such algorithms. For example, the work of the anthropologist Jan Czekanowski, in 1909, contains many of the ideas required to find minimum spanning trees. In 1926, Otakar Boruvka described methods for constructing minimum spanning trees in work relating to the construction of electric power networks.

(a)

Choice	Edge	Weight
1	$\{c, d\}$	1
2	$\{k, l\}$	1
3	$\{b, f\}$	1
4	$\{c, g\}$	2
5	$\{a, b\}$	2
6	$\{f, j\}$	2
7	$\{b, c\}$	3
8	$\{j, k\}$	3
9	$\{g, h\}$	3
10	$\{i, j\}$	3
11	$\{a, e\}$	3
	Total:	24

(b)

FIGURE 5 A Minimum Spanning Tree Produced by Kruskal's Algorithm.

We will now prove that Prim's algorithm produces a minimum spanning tree of a connected weighted graph.

Proof: Let G be a connected weighted graph. Suppose that the successive edges chosen by Prim's algorithm are $e_1, e_2, \ldots, e_{n-1}$. Let S be the tree with $e_1, e_2, \ldots, e_{n-1}$ as its edges, and let S_k be the tree with $e_1, e_2, \ldots, e_k$ as its edges. Let T be a minimum spanning tree of G containing the edges $e_1, e_2, \ldots, e_k$, where k is the maximum integer with the property that a minimum spanning tree exists containing the first k edges chosen by Prim's algorithm. The theorem follows if we can show that $S = T$.

Suppose that $S \neq T$, so that $k < n - 1$. Consequently, T contains $e_1, e_2, \ldots, e_k$, but not e_{k+1}. Consider the graph made up of T together with e_{k+1}. Because this graph is connected and has n edges, too many edges to be a tree, it must contain a simple circuit. This simple circuit must contain e_{k+1} because there was no simple circuit in T. Furthermore, there must be an edge in the simple circuit that does not belong to S_{k+1} because S_{k+1} is a tree. By starting at an endpoint of e_{k+1} that is also an endpoint of one of the edges $e_1, \ldots, e_k$, and following the circuit until it reaches an edge not in S_{k+1}, we can find an edge e not in S_{k+1} that has an endpoint that is also an endpoint of one of the edges $e_1, e_2, \ldots, e_k$. By deleting e from T and adding e_{k+1}, we obtain a tree T' with $n - 1$ edges (it is a tree because it has no simple circuits). Note that the tree T' contains $e_1, e_2, \ldots, e_k, e_{k+1}$. Furthermore, because e_{k+1} was chosen by Prim's algorithm at the $(k + 1)$st step, and e was also available at that step, the weight of e_{k+1} is less than or equal to the weight of e. From this observation it follows that T' is also a minimum spanning tree, because the sum of the weights of its edges does not exceed the sum of the weights of the edges of T. This contradicts the choice of k as the maximum integer such that a minimum spanning tree exists containing $e_1, \ldots, e_k$. Hence, $k = n - 1$, and $S = T$. It follows that Prim's algorithm produces a minimum spanning tree. ◁

It can be shown (see [CoLeRiSt01]) that to find a minimum spanning tree of a graph with e edges and v vertices, Kruskal's algorithm can be carried out using $O(e \log e)$ operations and Prim's algorithm can be carried out using $O(e \log v)$ operations. Consequently, it is preferable to use Kruskal's algorithm for graphs that are **sparse,** that is, where e is very small compared to $C(v, 2) = v(v - 1)/2$, the total number of possible edges in an undirected graph with v vertices. Otherwise, there is little difference in the complexity of these two algorithms.

Exercises

1. The roads represented by this graph are all unpaved. The lengths of the roads between pairs of towns are represented by edge weights. Which roads should be paved so that there is a path of paved roads between each pair of towns so that a minimum road length is paved? (*Note:* These towns are in Nevada.)

In Exercises 2–4 use Prim's algorithm to find a minimum spanning tree for the given weighted graph.

2.

3.

4.

5. Use Kruskal's algorithm to design the communications network described at the beginning of the section.
6. Use Kruskal's algorithm to find a minimum spanning tree for the weighted graph in Exercise 2.
7. Use Kruskal's algorithm to find a minimum spanning tree for the weighted graph in Exercise 3.
8. Use Kruskal's algorithm to find a minimum spanning tree for the weighted graph in Exercise 4.
9. Find a connected weighted simple graph with the fewest edges possible that has more than one minimum spanning tree.
10. A **minimum spanning forest** in a weighted graph is a spanning forest with minimal weight. Explain how Prim's and Kruskal's algorithms can be adapted to construct minimum spanning forests.

A **maximum spanning tree** of a connected weighted undirected graph is a spanning tree with the largest possible weight.

11. Devise an algorithm similar to Prim's algorithm for constructing a maximum spanning tree of a connected weighted graph.
12. Devise an algorithm similar to Kruskal's algorithm for constructing a maximum spanning tree of a connected weighted graph.
13. Find a maximum spanning tree for the weighted graph in Exercise 2.
14. Find a maximum spanning tree for the weighted graph in Exercise 3.
15. Find a maximum spanning tree for the weighted graph in Exercise 4.
16. Find the second least expensive communications network connecting the five computer centers in the problem posed at the beginning of the section.

*17. Devise an algorithm for finding the second shortest spanning tree in a connected weighted graph.

*18. Show that an edge with smallest weight in a connected weighted graph must be part of any minimum spanning tree.

19. Show that there is a unique minimum spanning tree in a connected weighted graph if the weights of the edges are all different.
20. Suppose that the computer network connecting the cities in Figure 1 must contain a direct link between New York and Denver. What other links should be included so that there is a link between every two computer centers and the cost is minimized?
21. Find a spanning tree with minimal total weight containing the edges $\{e, i\}$ and $\{g, k\}$ in the weighted graph in Figure 3.
22. Describe an algorithm for finding a spanning tree with minimal weight containing a specified set of edges in a connected weighted undirected simple graph.

23. Express the algorithm devised in Exercise 22 in pseudocode.

Sollin's algorithm produces a minimum spanning tree from a connected weighted simple graph $G = (V, E)$ by successively adding groups of edges. Suppose that the vertices in V are ordered. This produces an ordering of the edges where $\{u_0, v_0\}$ precedes $\{u_1, v_1\}$ if u_0 precedes u_1 or if $u_0 = u_1$ and v_0 precedes v_1. The algorithm begins by simultaneously choosing the edge of least weight incident to each vertex. The first edge in the ordering is taken in the case of ties. This produces a graph with no simple circuits, that is, a forest of trees (Exercise 24 asks for a proof of this fact). Next, simultaneously choose for each tree in the forest the shortest edge between a vertex in this tree and a vertex in a different tree. Again the first edge in the ordering is chosen in the case of ties. (This produces a graph with no simple circuits containing fewer trees than were present before this step; see Exercise 24.) Continue the process of simultaneously adding edges connecting trees until $n - 1$ edges have been chosen. At this stage a minimum spanning tree has been constructed.

*24. Show that the addition of edges at each stage of Sollin's algorithm produces a forest.

25. Use Sollin's algorithm to produce a minimum spanning tree for the weighted graph shown in
   a) Figure 1.
   b) Figure 3.

*26. Express Sollin's algorithm in pseudocode.

**27. Prove that Sollin's algorithm produces a minimum spanning tree in a connected undirected weighted graph.

*28. Show that the first step of Sollin's algorithm produces a forest containing at least $\lceil n/2 \rceil$ edges.

*29. Show that if there are r trees in the forest at some intermediate step of Sollin's algorithm, then at least $\lceil r/2 \rceil$ edges are added by the next iteration of the algorithm.

*30. Show that no more than $\lfloor n/2^k \rfloor$ trees remain after the first step of Sollin's algorithm has been carried out and the second step of the algorithm has been carried out $k - 1$ times.

*31. Show that Sollin's algorithm requires at most $\log n$ iterations to produce a minimum spanning tree from a connected undirected weighted graph with n vertices.

32. Prove that Kruskal's algorithm produces minimum spanning trees.

Key Terms and Results

TERMS

tree: a connected undirected graph with no simple circuits

forest: an undirected graph with no simple circuits

rooted tree: a directed graph with a specified vertex, called the root, such that there is a unique path to any other vertex from this root

subtree: a subgraph of a tree that is also a tree

parent of v in a rooted tree: the vertex u such that (u, v) is an edge of the rooted tree

child of a vertex v in a rooted tree: any vertex with v as its parent

sibling of a vertex v in a rooted tree: a vertex with the same parent as v

ancestor of a vertex v in a rooted tree: any vertex on the path from the root to v

descendant of a vertex v in a rooted tree: any vertex that has v as an ancestor

internal vertex: a vertex that has children

leaf: a vertex with no children

level of a vertex: the length of the path from the root to this vertex

height of a tree: the largest level of the vertices of a tree

m-ary tree: a tree with the property that every internal vertex has no more than m children

full m-ary tree: a tree with the property that every internal vertex has exactly m children

binary tree: an m-ary tree with $m = 2$ (each child may be designated as a left or a right child of its parent)

ordered tree: a tree in which the children of each internal vertex are linearly ordered

balanced tree: a tree in which every leaf is at level h or $h - 1$, where h is the height of the tree

binary search tree: a binary tree in which the vertices are labeled with items so that a label of a vertex is greater than the labels of all vertices in the left subtree of this vertex and is less than the labels of all vertices in the right subtree of this vertex

decision tree: a rooted tree where each vertex represents a possible outcome of a decision and the leaves represent the possible solutions

prefix code: a code that has the property that the code of a character is never a prefix of the code of another character

minmax strategy: the strategy where the first player and second player move to positions represented by a child with maximum and minimum value, respectively

value of a vertex in a game tree: for a leaf, the payoff to the first player when the game terminates in the position represented by this leaf; for an internal vertex, the maximum or minimum of the values of its children, for an internal vertex at an even or odd level, respectively

tree traversal: a listing of the vertices of a tree

preorder traversal: a listing of the vertices of an ordered rooted tree defined recursively—the root is listed, followed by the first subtree, followed by the other subtrees in the order they occur from left to right

inorder traversal: a listing of the vertices of an ordered rooted tree defined recursively—the first subtree is listed, followed by the root, followed by the other subtrees in the order they occur from left to right

postorder traversal: a listing of the vertices of an ordered rooted tree defined recursively—the subtrees are listed in

the order they occur from left to right, followed by the root

infix notation: the form of an expression (including a full set of parentheses) obtained from an inorder traversal of the binary tree representing this expression

prefix (or Polish) notation: the form of an expression obtained from a preorder traversal of the tree representing this expression

postfix (or reverse Polish) notation: the form of an expression obtained from a postorder traversal of the tree representing this expression

spanning tree: a tree containing all vertices of a graph

minimum spanning tree: a spanning tree with smallest possible sum of weights of its edges

greedy algorithm: an algorithm that optimizes by making the optimal choice at each step

RESULTS

A graph is a tree if and only if there is a unique simple path between any of its vertices.

A tree with n vertices has $n - 1$ edges.

A full m-ary tree with i internal vertices has $mi + 1$ vertices.

The relationships among the numbers of vertices, leaves, and internal vertices in a full m-ary tree (see Theorem 4 in Section 10.1)

There are at most m^h leaves in an m-ary tree of height h.

If an m-ary tree has l leaves, its height h is at least $\lceil \log_m l \rceil$. If the tree is also full and balanced, then its height is $\lceil \log_m l \rceil$.

Huffman coding: a procedure for constructing an optimal binary code for a set of symbols, given the frequencies of these symbols

depth-first search, or backtracking: a procedure for constructing a spanning tree by adding edges that form a path until this is not possible, and then moving back up the path until a vertex is found where a new path can be formed

breadth-first search: a procedure for constructing a spanning tree that successively adds all edges incident to the last set of edges added, unless a simple circuit is formed

Prim's algorithm: a procedure for producing a minimum spanning tree in a weighted graph that successively adds edges with minimal weight among all edges incident to a vertex already in the tree such that no edge produces a simple circuit when it is added

Kruskal's algorithm: a procedure for producing a minimum spanning tree in a weighted graph that successively adds edges of least weight that are not already in the tree such that no edge produces a simple circuit when it is added

Review Questions

1. **a)** Define a tree. **b)** Define a forest.
2. Can there be two different simple paths between the vertices of a tree?
3. Give at least three examples of how trees are used in modeling.
4. **a)** Define a rooted tree and the root of such a tree.
   b) Define the parent of a vertex and a child of a vertex in a rooted tree.
   c) What are an internal vertex, a leaf, and a subtree in a rooted tree?
   d) Draw a rooted tree with at least 10 vertices, where the degree of each vertex does not exceed 3. Identify the root, the parent of each vertex, the children of each vertex, the internal vertices, and the leaves.
5. **a)** How many edges does a tree with n vertices have?
   b) What do you need to know to determine the number of edges in a forest with n vertices?
6. **a)** Define a full m-ary tree.
   b) How many vertices does a full m-ary tree have if it has i internal vertices? How many leaves does the tree have?
7. **a)** What is the height of a rooted tree?
   b) What is a balanced tree?
   c) How many leaves can an m-ary tree of height h have?
8. **a)** What is a binary search tree?
   b) Describe an algorithm for constructing a binary search tree.
   c) Form a binary search tree for the words *vireo, warbler, egret, grosbeak, nuthatch,* and *kingfisher.*
9. **a)** What is a prefix code?
   b) How can a prefix code be represented by a binary tree?
10. **a)** Define preorder, inorder, and postorder tree traversal.
    b) Give an example of preorder, postorder, and inorder traversal of a binary tree of your choice with at least 12 vertices.
11. **a)** Explain how to use preorder, inorder, and postorder traversals to find the prefix, infix, and postfix forms of an arithmetic expression.
    b) Draw the ordered rooted tree that represents $((x - 3) + ((x/4) + (x - y) \uparrow 3))$.
    c) Find the prefix and postfix forms of the expression in part (b).
12. Show that the number of comparisons used by a sorting algorithm is at least $\lceil \log n! \rceil$.
13. **a)** Describe the Huffman coding algorithm for constructing an optimal code for a set of symbols, given the frequency of these symbols.
    b) Use Huffman coding to find an optimal code for these symbols and frequencies: A: 0.2, B: 0.1, C: 0.3, D: 0.4.
14. Draw the game tree for nim if the starting position consists of two piles with one and four stones, respectively. Who wins the game if both players follow an optimal strategy?
15. **a)** What is a spanning tree of a simple graph?

b) Which simple graphs have spanning trees?
c) Describe at least two different applications that require that a spanning tree of a simple graph be found.

16. a) Describe two different algorithms for finding a spanning tree in a simple graph.
b) Illustrate how the two algorithms you described in (a) can be used to find the spanning tree of a simple graph, using a graph of your choice with at least eight vertices and 15 edges.

17. a) Explain how backtracking can be used to determine whether a simple graph can be colored using n colors.
b) Show, with an example, how backtracking can be used to show that a graph with a chromatic number equal to 4 cannot be colored with three colors, but can be colored with four colors.

18. a) What is a minimum spanning tree of a connected weighted graph?
b) Describe at least two different applications that require that a minimum spanning tree of a connected weighted graph be found.

19. a) Describe Kruskal's algorithm and Prim's algorithm for finding minimum spanning trees.
b) Illustrate how Kruskal's algorithm and Prim's algorithm are used to find a minimum spanning tree, using a weighted graph with at least eight vertices and 15 edges.

Supplementary Exercises

*1. Show that a simple graph is a tree if and only if it contains no simple circuits and the addition of an edge connecting two nonadjacent vertices produces a new graph that has exactly one simple circuit (where circuits that contain the same edges are not considered different).

*2. How many nonisomorphic rooted trees are there with six vertices?

3. Show that every tree with at least one edge must have at least two pendant vertices.

4. Show that a tree with n vertices that has $n - 1$ pendant vertices must be isomorphic to $K_{1,n-1}$.

5. What is the sum of the degrees of the vertices of a tree with n vertices?

*6. Suppose that $d_1, d_2, \ldots, d_n$ are n positive integers with sum $2n - 2$. Show that there is a tree that has n vertices such that the degrees of these vertices are $d_1, d_2, \ldots, d_n$.

7. Show that every tree is a planar graph.

8. Show that every tree is bipartite.

9. Show that every forest can be colored using two colors.

A **B-tree of degree k** is a rooted tree such that all its leaves are at the same level, its root has at least two and at most k children unless it is a leaf, and every internal vertex other than the root has at least $\lceil k/2 \rceil$, but no more than k, children. Computer files can be accessed efficiently when B-trees are used to represent them.

10. Draw three different B-trees of degree 3 with height 4.

11. Give an upper bound and a lower bound for the number of leaves in a B-tree of degree k with height h.

12. Give an upper bound and a lower bound for the height of a B-tree of degree k with n leaves.

The **binomial trees** B_i, $i = 0, 1, 2, \ldots$, are ordered rooted trees defined recursively:

Basis step: The binomial tree B_0 is the tree with a single vertex.

Recursive step: Let k be a nonnegative integer. To construct the binomial tree B_{k+1}, add a copy of B_k to a second copy of B_k by adding an edge that makes the root of the first copy of B_k the leftmost child of the second copy of B_k.

13. Draw B_k for $k = 0, 1, 2, 3, 4$.

14. How many vertices does B_k have? Prove that your answer is correct.

15. Find the height of B_k. Prove that your answer is correct.

16. How many vertices are there in B_k at depth j, where $0 \le j \le k$? Justify your answer.

17. What is the degree of the root of B_k? Prove that your answer is correct.

18. Show that the vertex of largest degree in B_k is the root.

A rooted tree T is called an **S_k-tree** if it satisfies this recursive definition. It is an S_0-tree if it has one vertex. For $k > 0$, T is an S_k-tree if it can be built from two S_{k-1}-trees by making the root of one the root of the S_k-tree and making the root of the other the child of the root of the first S_k-tree.

19. Draw an S_k-tree for $k = 0, 1, 2, 3, 4$.

20. Show that an S_k-tree has 2^k vertices and a unique vertex at level k. This vertex at level k is called the **handle.**

*21. Suppose that T is an S_k-tree with handle v. Show that T can be obtained from disjoint trees $T_0, T_1, \ldots, T_{k-1}$, where v is not in any of these trees, where T_i is an S_i-tree for $i = 0, 1, \ldots, k - 1$, by connecting v to r_0 and r_i to r_{i+1} for $i = 0, 1, \ldots, k - 2$.

The listing of the vertices of an ordered rooted tree in **level order** begins with the root, followed by the vertices at level 1 from left to right, followed by the vertices at level 2 from left to right, and so on.

22. List the vertices of the ordered rooted trees in Figures 3 and 9 of Section 10.3 in level order.

23. Devise an algorithm for listing the vertices of an ordered rooted tree in level order.

*24. Devise an algorithm for determining if a set of universal addresses can be the addresses of the leaves of a rooted tree.

25. Devise an algorithm for constructing a rooted tree from the universal addresses of its leaves.

A **cut set** of a graph is a set of edges such that the removal of these edges produces a subgraph with more connected components than in the original graph, but no proper subset of this set of edges has this property.

26. Show that a cut set of a graph must have at least one edge in common with any spanning tree of this graph.

A **cactus** is a connected graph in which no edge is in more than one simple circuit not passing through any vertex other than its initial vertex more than once or its initial vertex other than at its terminal vertex (where two circuits that contain the same edges are not considered different).

27. Which of these graphs are cacti?

a)

b)

c)

28. Is a tree necessarily a cactus?

29. Show that a cactus is formed if we add a circuit containing new edges beginning and ending at a vertex of a tree.

***30.** Show that if every circuit not passing through any vertex other than its initial vertex more than once in a connected graph contains an odd number of edges, then this graph must be a cactus.

A **degree-constrained spanning tree** of a simple graph G is a spanning tree with the property that the degree of a vertex in this tree cannot exceed some specified bound. Degree-constrained spanning trees are useful in models of transportation systems where the number of roads at an intersection is limited, models of communications networks where the number of links entering a node is limited, and so on.

In Exercises 31–33 find a degree-constrained spanning tree of the given graph where each vertex has degree less than or equal to 3, or show that such a spanning tree does not exist.

31.

32.

33.

34. Show that a degree-constrained spanning tree of a simple graph in which each vertex has degree not exceeding 2 consists of a single Hamilton path in the graph.

35. A tree with n vertices is called **graceful** if its vertices can be labeled with the integers $1, 2, \ldots, n$ such that the absolute values of the difference of the labels of adjacent vertices are all different. Show that these trees are graceful.

a)

b)

c)

d)

A **caterpillar** is a tree that contains a simple path such that every vertex not contained in this path is adjacent to a vertex in the path.

36. Which of the graphs in Exercise 35 are caterpillars?

37. How many nonisomorphic caterpillars are there with six vertices?

****38.** **a)** Prove or disprove that all trees whose edges form a single path are graceful.

b) Prove or disprove that all caterpillars are graceful.

9. Suppose that in a long bit string the frequency of occurrence of a 0 bit is 0.9 and the frequency of a 1 bit is 0.1 and bits occur independently.
   a) Construct a Huffman code for the four blocks of two bits, 00, 01, 10, and 11. What is the average number of bits required to encode a bit string using this code?
   b) Construct a Huffman code for the eight blocks of three bits. What is the average number of bits required to encode a bit string using this code?

0. Suppose that G is a directed graph with no circuits. Describe how depth-first search can be used to carry out a topological sort of the vertices of G.

1. Suppose that e is an edge in a weighted graph that is incident to a vertex v such that the weight of e does not exceed the weight of any other edge incident to v. Show that there exists a minimum spanning tree containing this edge.

2. Three couples arrive at the bank of a river. Each of the wives is jealous and does not trust her husband when he is with one of the other wives (and perhaps with other people), but not with her. How can six people cross to the other side of the river using a boat that can hold no more than two people so that no husband is alone with a woman other than his wife? Use a graph theory model.

*43. Show that if no two edges in a weighted graph have the same weight, then the edge with least weight incident to a vertex v is included in every minimum spanning tree.

44. Find a minimum spanning tree of each of these graphs where the degree of each vertex in the spanning tree does not exceed 2.

a)

b)

Computer Projects

Write programs with these input and output.

1. Given the adjacency matrix of an undirected simple graph, determine whether the graph is a tree.
2. Given the adjacency matrix of a rooted tree and a vertex in the tree, find the parent, children, ancestors, descendants, and level of this vertex.
3. Given the list of edges of a rooted tree and a vertex in the tree, find the parent, children, ancestors, descendants, and level of this vertex.
4. Given a list of items, construct a binary search tree containing these items.
5. Given a binary search tree and an item, locate or add this item to the binary search tree.
6. Given the ordered list of edges of an ordered rooted tree, find the universal addresses of its vertices.
7. Given the ordered list of edges of an ordered rooted tree, list its vertices in preorder, inorder, and postorder.
8. Given an arithmetic expression in prefix form, find its value.
9. Given an arithmetic expression in postfix form, find its value.
0. Given the frequency of symbols, use Huffman coding to find an optimal code for these symbols.
11. Given an initial position in the game of nim, determine an optimal strategy for the first player.
12. Given the adjacency matrix of a connected undirected simple graph, find a spanning tree for this graph using depth-first search.
13. Given the adjacency matrix of a connected undirected simple graph, find a spanning tree for this graph using breadth-first search.
14. Given a set of positive integers and a positive integer N, use backtracking to find a subset of these integers that have N as their sum.

*15. Given the adjacency matrix of an undirected simple graph, use backtracking to color the graph with three colors, if this is possible.

*16. Given a positive integer n, solve the n-queens problem using backtracking.

17. Given the list of edges and their weights of a weighted undirected connected graph, use Prim's algorithm to find a minimum spanning tree of this graph.
18. Given the list of edges and their weights of a weighted undirected connected graph, use Kruskal's algorithm to find a minimum spanning tree of this graph.

Computations and Explorations

Use a computational program or programs you have written to do these exercises.

1. Display all trees with six vertices.
2. Display a full set of nonisomorphic trees with seven vertices.
*3. Construct a Huffman code for the symbols with ASCII codes given the frequency of their occurrence in representative input.
4. Compute the number of different spanning trees of K_n for $n = 1, 2, 3, 4, 5, 6$. Conjecture a formula for the number of such spanning trees whenever n is a positive integer.
5. Compare the number of comparisons needed to sort lists of n elements for $n = 100$, 1000, and 10,000, where the elements are randomly selected positive integers, using the selection sort, the insertion sort, the merge sort, and the quick sort.
6. Compute the number of different ways n queens can be arranged on an $n \times n$ chessboard so that no two queens can attack each other for all positive integers n not exceeding 10.
*7. Find a minimum spanning tree of the graph that connects the capital cities of the 50 states in the United States to each other where the weight of each edge is the distance between the cities.
8. Draw the complete game tree for a game of checkers on a 4×4 board.

Writing Projects

Respond to these with essays using outside sources.

1. Explain how Cayley used trees to enumerate the number of certain types of hydrocarbons.
2. Define *AVL-trees* (sometimes also known as *height-balanced trees*). Describe how and why AVL-trees are used in a variety of different algorithms.
3. Define *quad trees* and explain how images can be represented using them. Describe how images can be rotated, scaled, and translated by manipulating the corresponding quad tree.
4. Define a *heap* and explain how trees can be turned into heaps. Why are heaps useful in sorting?
5. Describe dynamic algorithms for data compression based on letter frequencies as they change as characters are successively read, such as adaptive Huffman coding.
6. Explain how *alpha-beta pruning* can be used to simplify the computation of the value of a game tree.
7. Describe the techniques used by chess-playing programs such as Deep Blue.
8. Define the type of graph known as a *mesh of trees*. Explain how this graph is used in applications to very large system integration and parallel computing.
9. Discuss the algorithms used in IP multicasting to avoid loops between routers.
10. Describe an algorithm based on depth-first search for finding the articulation points of a graph.
11. Describe an algorithm based on depth-first search to find the strongly connected components of a directed graph.
12. Describe the search techniques used by the crawlers and spiders in different search engines on the Web.
13. Describe an algorithm for finding the minimum spanning tree of a graph such that the maximum degree of any vertex in the spanning tree does not exceed a fixed constant k.
14. Compare and contrast some of the most important sorting algorithms in terms of their complexity and when they are used.
15. Discuss the history and origins of algorithms for constructing minimum spanning trees.
16. Describe algorithms for producing random trees.

CHAPTER

8 Relations

8.1 Relations and Their Properties

8.2 n-ary Relations and Their Applications

8.3 Representing Relations

8.4 Closures of Relations

8.5 Equivalence Relations

8.6 Partial Orderings

Relationships between elements of sets occur in many contexts. Every day we deal with relationships such as those between a business and its telephone number, an employee and his or her salary, a person and a relative, and so on. In mathematics we study relationships such as those between a positive integer and one that it divides, an integer and one that it is congruent to modulo 5, a real number and one that is larger than it, a real number x and the value $f(x)$ where f is a function, and so on. Relationships such as that between a program and a variable it uses and that between a computer language and a valid statement in this language often arise in computer science.

Relationships between elements of sets are represented using the structure called a relation, which is just a subset of the Cartesian product of the sets. Relations can be used to solve problems such as determining which pairs of cities are linked by airline flights in a network, finding a viable order for the different phases of a complicated project, or producing a useful way to store information in computer databases.

In some computer languages, only the first 31 characters of the name of a variable matter. The relation consisting of ordered pairs of strings where the first string has the same initial 31 characters as the second string is an example of a special type of relation, known as an equivalence relation. Equivalence relations arise throughout mathematics and computer science. We will study equivalence relations, and other special types of relations, in this chapter.

8.1 Relations and Their Properties

Introduction

The most direct way to express a relationship between elements of two sets is to use ordered pairs made up of two related elements. For this reason, sets of ordered pairs are called binary relations. In this section we introduce the basic terminology used to describe binary relations. Later in this chapter we will use relations to solve problems involving communications networks, project scheduling, and identifying elements in sets with common properties.

DEFINITION 1 Let A and B be sets. A *binary relation from A to B* is a subset of $A \times B$.

In other words, a binary relation from A to B is a set R of ordered pairs where the first element of each ordered pair comes from A and the second element comes from B. We use the notation $a\,R\,b$ to denote that $(a, b) \in R$ and $a \not R\, b$ to denote that $(a, b) \notin R$. Moreover, when (a, b) belongs to R, a is said to be **related to** b by R.

Binary relations represent relationships between the elements of two sets. We will introduce n-ary relations, which express relationships among elements of more than two sets, later in this chapter. We will omit the word *binary* when there is no danger of confusion.

Examples 1–3 illustrate the notion of a relation.

EXAMPLE 1 Let A be the set of students in your school, and let B be the set of courses. Let R be the relation that consists of those pairs (a, b), where a is a student enrolled in course b. For instance, if

R	a	b
0	×	×
1	×	
2		×

FIGURE 1 Displaying the Ordered Pairs in the Relation R from Example 3.

Jason Goodfriend and Deborah Sherman are enrolled in CS518, the pairs (Jason Goodfriend, CS518) and (Deborah Sherman, CS518) belong to R. If Jason Goodfriend is also enrolled in CS510, then the pair (Jason Goodfriend, CS510) is also in R. However, if Deborah Sherman is not enrolled in CS510, then the pair (Deborah Sherman, CS510) is not in R.

Note that if a student is not currently enrolled in any courses there will be no pairs in R that have this student as the first element. Similarly, if a course is not currently being offered there will be no pairs in R that have this course as their second element. ◀

EXAMPLE 2 Let A be the set of all cities, and let B be the set of the 50 states in the United States of America. Define the relation R by specifying that (a, b) belongs to R if city a is in state b. For instance, (Boulder, Colorado), (Bangor, Maine), (Ann Arbor, Michigan), (Middletown, New Jersey), (Middletown, New York), (Cupertino, California), and (Red Bank, New Jersey) are in R. ◀

EXAMPLE 3 Let $A = \{0, 1, 2\}$ and $B = \{a, b\}$. Then $\{(0, a), (0, b), (1, a), (2, b)\}$ is a relation from A to B. This means, for instance, that $0\,R\,a$, but that $1 \not R\, b$. Relations can be represented graphically, as shown in Figure 1, using arrows to represent ordered pairs. Another way to represent this relation is to use a table, which is also done in Figure 1. We will discuss representations of relations in more detail in Section 8.3. ◀

Functions as Relations

Recall that a function f from a set A to a set B (as defined in Section 2.3) assigns exactly one element of B to each element of A. The graph of f is the set of ordered pairs (a, b) such that $b = f(a)$. Because the graph of f is a subset of $A \times B$, it is a relation from A to B. Moreover, the graph of a function has the property that every element of A is the first element of exactly one ordered pair of the graph.

Conversely, if R is a relation from A to B such that every element in A is the first element of exactly one ordered pair of R, then a function can be defined with R as its graph. This can be done by assigning to an element a of A the unique element $b \in B$ such that $(a, b) \in R$. (Note that the relation R in Example 2 is not the graph of a function because Middletown occurs more than once as the first element of an ordered pair in R.)

A relation can be used to express a one-to-many relationship between the elements of the sets A and B (as in Example 2), where an element of A may be related to more than one element of B. A function represents a relation where exactly one element of B is related to each element of A.

Relations are a generalization of functions; they can be used to express a much wider class of relationships between sets.

Relations on a Set

Relations from a set A to itself are of special interest.

DEFINITION 2 A *relation on the set* A is a relation from A to A.

In other words, a relation on a set A is a subset of $A \times A$.

EXAMPLE 4 Let A be the set $\{1, 2, 3, 4\}$. Which ordered pairs are in the relation $R = \{(a, b) \mid a \text{ divides } b\}$?

Solution: Because (a, b) is in R if and only if a and b are positive integers not exceeding 4 such that a divides b, we see that

$$R = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 2), (2, 4), (3, 3), (4, 4)\}.$$

The pairs in this relation are displayed both graphically and in tabular form in Figure 2. ◀

Next, some examples of relations on the set of integers will be given in Example 5.

EXAMPLE 5 Consider these relations on the set of integers:

$$\begin{aligned}
R_1 &= \{(a, b) \mid a \leq b\},\\
R_2 &= \{(a, b) \mid a > b\},\\
R_3 &= \{(a, b) \mid a = b \text{ or } a = -b\},\\
R_4 &= \{(a, b) \mid a = b\},\\
R_5 &= \{(a, b) \mid a = b + 1\},\\
R_6 &= \{(a, b) \mid a + b \leq 3\}.
\end{aligned}$$

Which of these relations contain each of the pairs $(1, 1)$, $(1, 2)$, $(2, 1)$, $(1, -1)$, and $(2, 2)$?

Remark: Unlike the relations in Examples 1–4, these are relations on an infinite set.

Solution: The pair $(1, 1)$ is in R_1, R_3, R_4, and R_6; $(1, 2)$ is in R_1 and R_6; $(2, 1)$ is in R_2, R_5, and R_6; $(1, -1)$ is in R_2, R_3, and R_6; and finally, $(2, 2)$ is in R_1, R_3, and R_4. ◀

R	1	2	3	4
1	×	×	×	×
2		×		×
3			×	
4				×

FIGURE 2 Displaying the Ordered Pairs in the Relation R from Example 4.

It is not hard to determine the number of relations on a finite set, because a relation on a set A is simply a subset of $A \times A$.

EXAMPLE 6 How many relations are there on a set with n elements?

Solution: A relation on a set A is a subset of $A \times A$. Because $A \times A$ has n^2 elements when A has n elements, and a set with m elements has 2^m subsets, there are 2^{n^2} subsets of $A \times A$. Thus, there are 2^{n^2} relations on a set with n elements. For example, there are $2^{3^2} = 2^9 = 512$ relations on the set $\{a, b, c\}$. ◀

Properties of Relations

There are several properties that are used to classify relations on a set. We will introduce the most important of these here.

In some relations an element is always related to itself. For instance, let R be the relation on the set of all people consisting of pairs (x, y) where x and y have the same mother and the same father. Then xRx for every person x.

DEFINITION 3 A relation R on a set A is called *reflexive* if $(a, a) \in R$ for every element $a \in A$.

Remark: Using quantifiers we see that the relation R on the set A is reflexive if $\forall a((a, a) \in R)$, where the universe of discourse is the set of all elements in A.

We see that a relation on A is reflexive if every element of A is related to itself. Examples 7–9 illustrate the concept of a reflexive relation.

EXAMPLE 7 Consider the following relations on $\{1, 2, 3, 4\}$:

$$R_1 = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 4), (4, 1), (4, 4)\},$$
$$R_2 = \{(1, 1), (1, 2), (2, 1)\},$$
$$R_3 = \{(1, 1), (1, 2), (1, 4), (2, 1), (2, 2), (3, 3), (4, 1), (4, 4)\},$$
$$R_4 = \{(2, 1), (3, 1), (3, 2), (4, 1), (4, 2), (4, 3)\},$$
$$R_5 = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 2), (2, 3), (2, 4), (3, 3), (3, 4), (4, 4)\},$$
$$R_6 = \{(3, 4)\}.$$

Which of these relations are reflexive?

Solution: The relations R_3 and R_5 are reflexive because they both contain all pairs of the form (a, a), namely, $(1, 1)$, $(2, 2)$, $(3, 3)$, and $(4, 4)$. The other relations are not reflexive because they do not contain all of these ordered pairs. In particular, R_1, R_2, R_4, and R_6 are not reflexive because $(3, 3)$ is not in any of these relations. ◀

EXAMPLE 8 Which of the relations from Example 5 are reflexive?

Solution: The reflexive relations from Example 5 are R_1 (because $a \leq a$ for every integer a), R_3, and R_4. For each of the other relations in this example it is easy to find a pair of the form (a, a) that is not in the relation. (This is left as an exercise for the reader.) ◀

EXAMPLE 9 Is the "divides" relation on the set of positive integers reflexive?

Solution: Because $a \mid a$ whenever a is a positive integer, the "divides" relation is reflexive. (Note that if we replace the set of positive integers with the set of all integers the relation is not reflexive because 0 does not divide 0.) ◀

In some relations an element is related to a second element if and only if the second element is also related to the first element. The relation consisting of pairs (x, y), where x and y are students at your school with at least one common class has this property. Other relations have the property that if an element is related to a second element, then this second element is not related to the first. The relation consisting of the pairs (x, y), where x and y are students at your school, where x has a higher grade point average than y has this property.

DEFINITION 4 A relation R on a set A is called *symmetric* if $(b, a) \in R$ whenever $(a, b) \in R$, for all $a, b \in A$. A relation R on a set A such that for all $a, b \in A$, if $(a, b) \in R$ and $(b, a) \in R$, then $a = b$ is called *antisymmetric.*

Remark: Using quantifiers, we see that the relation R on the set A is symmetric if $\forall a \forall b((a, b) \in R \rightarrow (b, a) \in R)$. Similarly, the relation R on the set A is antisymmetric if $\forall a \forall b(((a, b) \in R \wedge (b, a) \in R) \rightarrow (a = b))$.

That is, a relation is symmetric if and only if a is related to b implies that b is related to a. A relation is antisymmetric if and only if there are no pairs of distinct elements a and b with a related to b and b related to a. That is, the only way to have a related to b and b related to a is for a and b to be the same element. The terms *symmetric* and *antisymmetric* are not opposites, because a relation can have both of these properties or may lack both of them (see Exercise 8 at the end of this section). A relation cannot be both symmetric and antisymmetric if it contains some pair of the form (a, b), where $a \neq b$.

Remark: Although relatively few of the 2^{n^2} relations on a set with n elements are symmetric or antisymmetric, as counting arguments can show, many important relations have one of these properties. (See Exercise 45.)

EXAMPLE 10 Which of the relations from Example 7 are symmetric and which are antisymmetric?

Extra Examples

Solution: The relations R_2 and R_3 are symmetric, because in each case (b, a) belongs to the relation whenever (a, b) does. For R_2, the only thing to check is that both $(2, 1)$ and $(1, 2)$ are in the relation. For R_3, it is necessary to check that both $(1, 2)$ and $(2, 1)$ belong to the relation, and $(1, 4)$ and $(4, 1)$ belong to the relation. The reader should verify that none of the other relations is symmetric. This is done by finding a pair (a, b) such that it is in the relation but (b, a) is not.

R_4, R_5, and R_6 are all antisymmetric. For each of these relations there is no pair of elements a and b with $a \neq b$ such that both (a, b) and (b, a) belong to the relation. The reader should verify that none of the other relations is antisymmetric. This is done by finding a pair (a, b) with $a \neq b$ such that (a, b) and (b, a) are both in the relation. ◀

EXAMPLE 11 Which of the relations from Example 5 are symmetric and which are antisymmetric?

Solution: The relations R_3, R_4, and R_6 are symmetric. R_3 is symmetric, for if $a = b$ or $a = -b$, then $b = a$ or $b = -a$. R_4 is symmetric because $a = b$ implies that $b = a$. R_6 is symmetric

because $a + b \leq 3$ implies that $b + a \leq 3$. The reader should verify that none of the other relations is symmetric.

The relations R_1, R_2, R_4, and R_5 are antisymmetric. R_1 is antisymmetric because the inequalities $a \leq b$ and $b \leq a$ imply that $a = b$. R_2 is antisymmetric because it is impossible for $a > b$ and $b > a$. R_4 is antisymmetric, because two elements are related with respect to R_4 if and only if they are equal. R_5 is antisymmetric because it is impossible that $a = b + 1$ and $b = a + 1$. The reader should verify that none of the other relations is antisymmetric. ◀

EXAMPLE 12 Is the "divides" relation on the set of positive integers symmetric? Is it antisymmetric?

Solution: This relation is not symmetric because $1 \mid 2$, but $2 \nmid 1$. It is antisymmetric, for if a and b are positive integers with $a \mid b$ and $b \mid a$, then $a = b$ (the verification of this is left as an exercise for the reader). ◀

Let R be the relation consisting of all pairs (x, y) of students at your school, where x has taken more credits than y. Suppose that x is related to y and y is related to z. This means that x has taken more credits than y and y has taken more credits than z. We can conclude that x has taken more credits than z, so that x is related to z. What we have shown is that R has the transitive property, which is defined as follows.

DEFINITION 5 A relation R on a set A is called *transitive* if whenever $(a, b) \in R$ and $(b, c) \in R$, then $(a, c) \in R$, for all $a, b, c \in A$.

Remark: Using quantifiers we see that the relation R on a set A is transitive if we have $\forall a \forall b \forall c(((a, b) \in R \wedge (b, c) \in R) \rightarrow (a, c) \in R)$.

EXAMPLE 13 Which of the relations in Example 7 are transitive?

Extra Examples

Solution: R_4, R_5, and R_6 are transitive. For each of these relations, we can show that it is transitive by verifying that if (a, b) and (b, c) belong to this relation, then (a, c) also does. For instance, R_4 is transitive, because $(3, 2)$ and $(2, 1)$, $(4, 2)$ and $(2, 1)$, $(4, 3)$ and $(3, 1)$, and $(4, 3)$ and $(3, 2)$ are the only such sets of pairs, and $(3, 1)$, $(4, 1)$, and $(4, 2)$ belong to R_4. The reader should verify that R_5 and R_6 are transitive.

R_1 is not transitive because $(3, 4)$ and $(4, 1)$ belong to R_1, but $(3, 1)$ does not. R_2 is not transitive because $(2, 1)$ and $(1, 2)$ belong to R_2, but $(2, 2)$ does not. R_3 is not transitive because $(4, 1)$ and $(1, 2)$ belong to R_3, but $(4, 2)$ does not. ◀

EXAMPLE 14 Which of the relations in Example 5 are transitive?

Solution: The relations R_1, R_2, R_3, and R_4 are transitive. R_1 is transitive because $a \leq b$ and $b \leq c$ imply that $a \leq c$. R_2 is transitive because $a > b$ and $b > c$ imply that $a > c$. R_3 is transitive because $a = \pm b$ and $b = \pm c$ imply that $a = \pm c$. R_4 is clearly transitive, as the reader should verify. R_5 is not transitive because $(2, 1)$ and $(1, 0)$ belong to R_5, but $(2, 0)$ does not. R_6 is not transitive because $(2, 1)$ and $(1, 2)$ belong to R_6, but $(2, 2)$ does not. ◀

EXAMPLE 15 Is the "divides" relation on the set of positive integers transitive?

Solution: Suppose that a divides b and b divides c. Then there are positive integers k and l such that $b = ak$ and $c = bl$. Hence, $c = a(kl)$, so a divides c. It follows that this relation is transitive. ◀

We can use counting techniques to determine the number of relations with specific properties. Finding the number of relations with a particular property provides information about how common this property is in the set of all relations on a set with n elements.

EXAMPLE 16 How many reflexive relations are there on a set with n elements?

Solution: A relation R on a set A is a subset of $A \times A$. Consequently, a relation is determined by specifying whether each of the n^2 ordered pairs in $A \times A$ is in R. However, if R is reflexive, each of the n ordered pairs (a, a) for $a \in A$ must be in R. Each of the other $n(n-1)$ ordered pairs of the form (a, b), where $a \neq b$, may or may not be in R. Hence, by the product rule for counting, there are $2^{n(n-1)}$ reflexive relations [this is the number of ways to choose whether each element (a, b), with $a \neq b$, belongs to R]. ◀

The number of symmetric relations and the number of antisymmetric relations on a set with n elements can be found using reasoning similar to that in Example 16 (see Exercise 45 at the end of this section). Counting the transitive relations on a set with n elements is a problem beyond the scope of this book.

Combining Relations

Because relations from A to B are subsets of $A \times B$, two relations from A to B can be combined in any way two sets can be combined. Consider Examples 17–19.

EXAMPLE 17 Let $A = \{1, 2, 3\}$ and $B = \{1, 2, 3, 4\}$. The relations $R_1 = \{(1, 1), (2, 2), (3, 3)\}$ and $R_2 = \{(1, 1), (1, 2), (1, 3), (1, 4)\}$ can be combined to obtain

$$\begin{aligned}
R_1 \cup R_2 &= \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 2), (3, 3)\},\\
R_1 \cap R_2 &= \{(1, 1)\},\\
R_1 - R_2 &= \{(2, 2), (3, 3)\},\\
R_2 - R_1 &= \{(1, 2), (1, 3), (1, 4)\}.
\end{aligned}$$

◀

EXAMPLE 18 Let A and B be the set of all students and the set of all courses at a school, respectively. Suppose that R_1 consists of all ordered pairs (a, b), where a is a student who has taken course b, and R_2 consists of all ordered pairs (a, b), where a is a student who requires course b to graduate. What are the relations $R_1 \cup R_2$, $R_1 \cap R_2$, $R_1 \oplus R_2$, $R_1 - R_2$, and $R_2 - R_1$?

Solution: The relation $R_1 \cup R_2$ consists of all ordered pairs (a, b), where a is a student who either has taken course b or needs course b to graduate, and $R_1 \cap R_2$ is the set of all ordered pairs (a, b), where a is a student who has taken course b and needs this course to graduate. Also, $R_1 \oplus R_2$ consists of all ordered pairs (a, b), where student a has taken course b but does not need it to graduate or needs course b to graduate but has not taken it. $R_1 - R_2$ is the set of ordered pairs (a, b), where a has taken course b but does not need it to graduate; that is, b is an elective course that a has taken. $R_2 - R_1$ is the set of all ordered pairs (a, b), where b is a course that a needs to graduate but has not taken. ◀

EXAMPLE 19 Let R_1 be the "less than" relation on the set of real numbers and let R_2 be the "greater than" relation on the set of real numbers, that is, $R_1 = \{(x, y) \mid x < y\}$ and $R_2 = \{(x, y) \mid x > y\}$. What are $R_1 \cup R_2$, $R_1 \cap R_2$, $R_1 - R_2$, $R_2 - R_1$, and $R_1 \oplus R_2$?

Solution: We note that $(x, y) \in R_1 \cup R_2$ if and only if $(x, y) \in R_1$ or $(x, y) \in R_2$. Hence, $(x, y) \in R_1 \cup R_2$ if and only if $x < y$ or $x > y$. Because the condition $x < y$ or $x > y$ is the same as the condition $x \neq y$, it follows that $R_1 \cup R_2 = \{(x, y) \mid x \neq y\}$. In other words, the union of the "less than" relation and the "greater than" relation is the "not equals" relation.

Next, note that it is impossible for a pair (x, y) to belong to both R_1 and R_2 because it is impossible for $x < y$ and $x > y$. It follows that $R_1 \cap R_2 = \emptyset$. We also see that $R_1 - R_2 = R_1$, $R_2 - R_1 = R_2$, and $R_1 \oplus R_2 = R_1 \cup R_2 - R_1 \cap R_2 = \{(x, y) \mid x \neq y\}$. ◀

There is another way that relations are combined that is analogous to the composition of functions.

DEFINITION 6 Let R be a relation from a set A to a set B and S a relation from B to a set C. The *composite* of R and S is the relation consisting of ordered pairs (a, c), where $a \in A$, $c \in C$, and for which there exists an element $b \in B$ such that $(a, b) \in R$ and $(b, c) \in S$. We denote the composite of R and S by $S \circ R$.

Computing the composite of two relations requires that we find elements that are the second element of ordered pairs in the first relation and the first element of ordered pairs in the second relation, as Examples 20 and 21 illustrate.

EXAMPLE 20 What is the composite of the relations R and S, where R is the relation from $\{1, 2, 3\}$ to $\{1, 2, 3, 4\}$ with $R = \{(1, 1), (1, 4), (2, 3), (3, 1), (3, 4)\}$ and S is the relation from $\{1, 2, 3, 4\}$ to $\{0, 1, 2\}$ with $S = \{(1, 0), (2, 0), (3, 1), (3, 2), (4, 1)\}$?

Solution: $S \circ R$ is constructed using all ordered pairs in R and ordered pairs in S, where the second element of the ordered pair in R agrees with the first element of the ordered pair in S. For example, the ordered pairs $(2, 3)$ in R and $(3, 1)$ in S produce the ordered pair $(2, 1)$ in $S \circ R$. Computing all the ordered pairs in the composite, we find

$$S \circ R = \{(1, 0), (1, 1), (2, 1), (2, 2), (3, 0), (3, 1)\}.$$ ◀

EXAMPLE 21 **Composing the Parent Relation with Itself** Let R be the relation on the set of all people such that $(a, b) \in R$ if person a is a parent of person b. Then $(a, c) \in R \circ R$ if and only if there is a person b such that $(a, b) \in R$ and $(b, c) \in R$, that is, if and only if there is a person b such that a is a parent of b and b is a parent of c. In other words, $(a, c) \in R \circ R$ if and only if a is a grandparent of c. ◀

The powers of a relation R can be recursively defined from the definition of a composite of two relations.

DEFINITION 7 Let R be a relation on the set A. The powers R^n, $n = 1, 2, 3, \ldots$, are defined recursively by

$$R^1 = R \qquad \text{and} \qquad R^{n+1} = R^n \circ R.$$

The definition shows that $R^2 = R \circ R$, $R^3 = R^2 \circ R = (R \circ R) \circ R$, and so on.

EXAMPLE 22 Let $R = \{(1, 1), (2, 1), (3, 2), (4, 3)\}$. Find the powers R^n, $n = 2, 3, 4, \ldots$.

Solution: Because $R^2 = R \circ R$, we find that $R^2 = \{(1, 1), (2, 1), (3, 1), (4, 2)\}$. Furthermore, because $R^3 = R^2 \circ R$, $R^3 = \{(1, 1), (2, 1), (3, 1), (4, 1)\}$. Additional computation shows that R^4 is the same as R^3, so $R^4 = \{(1, 1), (2, 1), (3, 1), (4, 1)\}$. It also follows that $R^n = R^3$ for $n = 5, 6, 7, \ldots$. The reader should verify this. ◀

The following theorem shows that the powers of a transitive relation are subsets of this relation. It will be used in Section 8.4.

THEOREM 1 The relation R on a set A is transitive if and only if $R^n \subseteq R$ for $n = 1, 2, 3, \ldots$.

Proof: We first prove the "if" part of the theorem. We suppose that $R^n \subseteq R$ for $n = 1, 2, 3, \ldots$. In particular, $R^2 \subseteq R$. To see that this implies R is transitive, note that if $(a, b) \in R$ and $(b, c) \in R$, then by the definition of composition, $(a, c) \in R^2$. Because $R^2 \subseteq R$, this means that $(a, c) \in R$. Hence, R is transitive.

We will use mathematical induction to prove the only if part of the theorem. Note that this part of the theorem is trivially true for $n = 1$.

Assume that $R^n \subseteq R$, where n is a positive integer. This is the inductive hypothesis. To complete the inductive step we must show that this implies that R^{n+1} is also a subset of R. To show this, assume that $(a, b) \in R^{n+1}$. Then, because $R^{n+1} = R^n \circ R$, there is an element x with $x \in A$ such that $(a, x) \in R$ and $(x, b) \in R^n$. The inductive hypothesis, namely, that $R^n \subseteq R$, implies that $(x, b) \in R$. Furthermore, because R is transitive, and $(a, x) \in R$ and $(x, b) \in R$, it follows that $(a, b) \in R$. This shows that $R^{n+1} \subseteq R$, completing the proof. ◁

Exercises

1. List the ordered pairs in the relation R from $A = \{0, 1, 2, 3, 4\}$ to $B = \{0, 1, 2, 3\}$, where $(a, b) \in R$ if and only if
   a) $a = b$. **b)** $a + b = 4$.
   c) $a > b$. **d)** $a \mid b$.
   e) $\gcd(a, b) = 1$. **f)** $\operatorname{lcm}(a, b) = 2$.
2. **a)** List all the ordered pairs in the relation $R = \{(a, b) \mid a \text{ divides } b\}$ on the set $\{1, 2, 3, 4, 5, 6\}$.
   b) Display this relation graphically, as was done in Example 4.
   c) Display this relation in tabular form, as was done in Example 4.
3. For each of these relations on the set $\{1, 2, 3, 4\}$, decide whether it is reflexive, whether it is symmetric, whether it is antisymmetric, and whether it is transitive.
   a) $\{(2, 2), (2, 3), (2, 4), (3, 2), (3, 3), (3, 4)\}$
   b) $\{(1, 1), (1, 2), (2, 1), (2, 2), (3, 3), (4, 4)\}$
   c) $\{(2, 4), (4, 2)\}$
   d) $\{(1, 2), (2, 3), (3, 4)\}$
   e) $\{(1, 1), (2, 2), (3, 3), (4, 4)\}$
   f) $\{(1, 3), (1, 4), (2, 3), (2, 4), (3, 1), (3, 4)\}$
4. Determine whether the relation R on the set of all people is reflexive, symmetric, antisymmetric, and/or transitive, where $(a, b) \in R$ if and only if
   a) a is taller than b.
   b) a and b were born on the same day.
   c) a has the same first name as b.
   d) a and b have a common grandparent.
5. Determine whether the relation R on the set of all Web pages is reflexive, symmetric, antisymmetric, and/or transitive, where $(a, b) \in R$ if and only if
   a) everyone who has visited Web page a has also visited Web page b.
   b) there are no common links found on both Web page a and Web page b.
   c) there is at least one common link on Web page a and Web page b.
   d) there is a Web page that includes links to both Web page a and Web page b.
6. Determine whether the relation R on the set of all real numbers is reflexive, symmetric, antisymmetric, and/or transitive, where $(x, y) \in R$ if and only if

a) $x + y = 0$. b) $x = \pm y$.
c) $x - y$ is a rational number. d) $x = 2y$.
e) $xy \geq 0$. f) $xy = 0$.
g) $x = 1$. h) $x = 1$ or $y = 1$.

7. Determine whether the relation R on the set of all integers is reflexive, symmetric, antisymmetric, and/or transitive, where $(x, y) \in R$ if and only if
   a) $x \neq y$. b) $xy \geq 1$.
   c) $x = y + 1$ or $x = y - 1$.
   d) $x \equiv y \pmod 7$. e) x is a multiple of y.
   f) x and y are both negative or both nonnegative.
   g) $x = y^2$. h) $x \geq y^2$.
8. Give an example of a relation on a set that is
   a) symmetric and antisymmetric.
   b) neither symmetric nor antisymmetric.

A relation R on the set A is **irreflexive** if for every $a \in A$, $(a, a) \notin R$. That is, R is irreflexive if no element in A is related to itself.

9. Which relations in Exercise 3 are irreflexive?
10. Which relations in Exercise 4 are irreflexive?
11. Which relations in Exercise 5 are irreflexive?
12. Which relations in Exercise 6 are irreflexive?
13. Can a relation on a set be neither reflexive nor irreflexive?
14. Use quantifiers to express what it means for a relation to be irreflexive.
15. Give an example of an irreflexive relation on the set of all people.

A relation R is called **asymmetric** if $(a, b) \in R$ implies that $(b, a) \notin R$. Exercises 16–22 explore the notion of an asymmetric relation. Exercise 20 focuses on the difference between asymmetry and antisymmetry.

16. Which relations in Exercise 3 are asymmetric?
17. Which relations in Exercise 4 are asymmetric?
18. Which relations in Exercise 5 are asymmetric?
19. Which relations in Exercise 6 are asymmetric?
20. Must an asymmetric relation also be antisymmetric? Must an antisymmetric relation be asymmetric? Give reasons for your answers.
21. Use quantifiers to express what it means for a relation to be asymmetric.
22. Give an example of an asymmetric relation on the set of all people.
23. How many different relations are there from a set with m elements to a set with n elements?

☞ Let R be a relation from a set A to a set B. The **inverse relation** from B to A, denoted by R^{-1}, is the set of ordered pairs $\{(b, a) \mid (a, b) \in R\}$. The **complementary relation** $\overline{R}$ is the set of ordered pairs $\{(a, b) \mid (a, b) \notin R\}$.

24. Let R be the relation $R = \{(a, b) \mid a < b\}$ on the set of integers. Find
   a) R^{-1}. b) $\overline{R}$.
25. Let R be the relation $R = \{(a, b) \mid a \text{ divides } b\}$ on the set of positive integers. Find
   a) R^{-1}. b) $\overline{R}$.
26. Let R be the relation on the set of all states in the United States consisting of pairs (a, b) where state a borders state b. Find
   a) R^{-1}. b) $\overline{R}$.
27. Suppose that the function f from A to B is a one-to-one correspondence. Let R be the relation that equals the graph of f. That is, $R = \{(a, f(a)) \mid a \in A\}$. What is the inverse relation R^{-1}?
28. Let $R_1 = \{(1, 2), (2, 3), (3, 4)\}$ and $R_2 = \{(1, 1), (1, 2), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3), (3, 4)\}$ be relations from $\{1, 2, 3\}$ to $\{1, 2, 3, 4\}$. Find
   a) $R_1 \cup R_2$. b) $R_1 \cap R_2$.
   c) $R_1 - R_2$. d) $R_2 - R_1$.
29. Let A be the set of students at your school and B the set of books in the school library. Let R_1 and R_2 be the relations consisting of all ordered pairs (a, b), where student a is required to read book b in a course, and where student a has read book b, respectively. Describe the ordered pairs in each of these relations.
   a) $R_1 \cup R_2$ b) $R_1 \cap R_2$
   c) $R_1 \oplus R_2$ d) $R_1 - R_2$
   e) $R_2 - R_1$
30. Let R be the relation $\{(1, 2), (1, 3), (2, 3), (2, 4), (3, 1)\}$, and let S be the relation $\{(2, 1), (3, 1), (3, 2), (4, 2)\}$. Find $S \circ R$.
31. Let R be the relation on the set of people consisting of pairs (a, b), where a is a parent of b. Let S be the relation on the set of people consisting of pairs (a, b), where a and b are siblings (brothers or sisters). What are $S \circ R$ and $R \circ S$?

Exercises 32–35 deal with these relations on the set of real numbers:

$R_1 = \{(a, b) \in \mathbf{R}^2 \mid a > b\}$, the "greater than" relation,
$R_2 = \{(a, b) \in \mathbf{R}^2 \mid a \geq b\}$, the "greater than or equal to" relation,
$R_3 = \{(a, b) \in \mathbf{R}^2 \mid a < b\}$, the "less than" relation,
$R_4 = \{(a, b) \in \mathbf{R}^2 \mid a \leq b\}$, the "less than or equal to" relation,
$R_5 = \{(a, b) \in \mathbf{R}^2 \mid a = b\}$, the "equal to" relation,
$R_6 = \{(a, b) \in \mathbf{R}^2 \mid a \neq b\}$, the "unequal to" relation.

32. Find
   a) $R_1 \cup R_3$. b) $R_1 \cup R_5$.
   c) $R_2 \cap R_4$. d) $R_3 \cap R_5$.
   e) $R_1 - R_2$. f) $R_2 - R_1$.
   g) $R_1 \oplus R_3$. h) $R_2 \oplus R_4$.
33. Find
   a) $R_2 \cup R_4$. b) $R_3 \cup R_6$.
   c) $R_3 \cap R_6$. d) $R_4 \cap R_6$.
   e) $R_3 - R_6$. f) $R_6 - R_3$.
   g) $R_2 \oplus R_6$. h) $R_3 \oplus R_5$.

34. Find

a) $R_1 \circ R_1$. **b)** $R_1 \circ R_2$.
c) $R_1 \circ R_3$. **d)** $R_1 \circ R_4$.
e) $R_1 \circ R_5$. **f)** $R_1 \circ R_6$.
g) $R_2 \circ R_3$. **h)** $R_3 \circ R_3$.

35. Find

a) $R_2 \circ R_1$. **b)** $R_2 \circ R_2$.
c) $R_3 \circ R_5$. **d)** $R_4 \circ R_1$.
e) $R_5 \circ R_3$. **f)** $R_3 \circ R_6$.
g) $R_4 \circ R_6$. **h)** $R_6 \circ R_6$.

36. Let R be the parent relation on the set of all people (see Example 21). When is an ordered pair in the relation R^3?

37. Let R be the relation on the set of people with doctorates such that $(a, b) \in R$ if and only if a was the thesis advisor of b. When is an ordered pair (a, b) in R^2? When is an ordered pair (a, b) in R^n, when n is a positive integer? (Note that every person with a doctorate has a thesis advisor.)

38. Let R_1 and R_2 be the "divides" and "is a multiple of" relations on the set of all positive integers, respectively. That is, $R_1 = \{(a, b) \mid a \text{ divides } b\}$ and $R_2 = \{(a, b) \mid a \text{ is a multiple of } b\}$. Find

a) $R_1 \cup R_2$. **b)** $R_1 \cap R_2$.
c) $R_1 - R_2$. **d)** $R_2 - R_1$.
e) $R_1 \oplus R_2$.

39. Let R_1 and R_2 be the "congruent modulo 3" and the "congruent modulo 4" relations, respectively, on the set of integers. That is, $R_1 = \{(a, b) \mid a \equiv b \pmod 3\}$ and $R_2 = \{(a, b) \mid a \equiv b \pmod 4\}$. Find

a) $R_1 \cup R_2$. **b)** $R_1 \cap R_2$.
c) $R_1 - R_2$. **d)** $R_2 - R_1$.
e) $R_1 \oplus R_2$.

40. List the 16 different relations on the set $\{0, 1\}$.

41. How many of the 16 different relations on $\{0, 1\}$ contain the pair $(0, 1)$?

42. Which of the 16 relations on $\{0, 1\}$, which you listed in Exercise 40, are

a) reflexive? **b)** irreflexive?
c) symmetric? **d)** antisymmetric?
e) asymmetric? **f)** transitive?

43. **a)** How many relations are there on the set $\{a, b, c, d\}$?
b) How many relations are there on the set $\{a, b, c, d\}$ that contain the pair (a, a)?

44. Let S be a set with n elements and let a and b be distinct elements of S. How many relations are there on S such that

a) $(a, b) \in S$? **b)** $(a, b) \notin S$?
c) there are no ordered pairs in the relation that have a as their first element?
d) there is at least one ordered pair in the relation that has a as its first element?
e) there are no ordered pairs in the relation that have a as their first element and there are no ordered pairs in the relation that have b as their second element?
f) there is at least one ordered pair in the relation that either has a as its first element or has b as its second element?

***45.** How many relations are there on a set with n elements that are

a) symmetric? **b)** antisymmetric?
c) asymmetric? **d)** irreflexive?
e) reflexive and symmetric?
f) neither reflexive nor irreflexive?

***46.** How many transitive relations are there on a set with n elements if

a) $n = 1$? **b)** $n = 2$? **c)** $n = 3$?

47. Find the error in the "proof" of the following "theorem."
"Theorem": Let R be a relation on a set A that is symmetric and transitive. Then R is reflexive.

"Proof": Let $a \in A$. Take an element $b \in A$ such that $(a, b) \in R$. Because R is symmetric, we also have $(b, a) \in R$. Now using the transitive property, we can conclude that $(a, a) \in R$ because $(a, b) \in R$ and $(b, a) \in R$.

48. Suppose that R and S are reflexive relations on a set A. Prove or disprove each of these statements.

a) $R \cup S$ is reflexive.
b) $R \cap S$ is reflexive.
c) $R \oplus S$ is irreflexive.
d) $R - S$ is irreflexive.
e) $S \circ R$ is reflexive.

49. Show that the relation R on a set A is symmetric if and only if $R = R^{-1}$, where R^{-1} is the inverse relation.

50. Show that the relation R on a set A is antisymmetric if and only if $R \cap R^{-1}$ is a subset of the diagonal relation $\Delta = \{(a, a) \mid a \in A\}$.

51. Show that the relation R on a set A is reflexive if and only if the inverse relation R^{-1} is reflexive.

52. Show that the relation R on a set A is reflexive if and only if the complementary relation $\overline{R}$ is irreflexive.

53. Let R be a relation that is reflexive and transitive. Prove that $R^n = R$ for all positive integers n.

54. Let R be the relation on the set $\{1, 2, 3, 4, 5\}$ containing the ordered pairs $(1, 1)$, $(1, 2)$, $(1, 3)$, $(2, 3)$, $(2, 4)$, $(3, 1)$, $(3, 4)$, $(3, 5)$, $(4, 2)$, $(4, 5)$, $(5, 1)$, $(5, 2)$, and $(5, 4)$. Find

a) R^2. **b)** R^3. **c)** R^4. **d)** R^5.

55. Let R be a reflexive relation on a set A. Show that R^n is reflexive for all positive integers n.

***56.** Let R be a symmetric relation. Show that R^n is symmetric for all positive integers n.

57. Suppose that the relation R is irreflexive. Is R^2 necessarily irreflexive? Give a reason for your answer.

8.2 n-ary Relations and Their Applications

Introduction

Relationships among elements of more than two sets often arise. For instance, there is a relationship involving the name of a student, the student's major, and the student's grade point average. Similarly, there is a relationship involving the airline, flight number, starting point, destination, departure time, and arrival time of a flight. An example of such a relationship in mathematics involves three integers, where the first integer is larger than the second integer, which is larger than the third. Another example is the betweenness relationship involving points on a line, such that three points are related when the second point is between the first and the third.

We will study relationships among elements from more than two sets in this section. These relationships are called ***n*-ary relations.** These relations are used to represent computer databases. These representations help us answer queries about the information stored in databases, such as: Which flights land at O'Hare Airport between 3 A.M. and 4 A.M.? Which students at your school are sophomores majoring in mathematics or computer science and have greater than a 3.0 average? Which employees of a company have worked for the company less than 5 years and make more than $50,000?

n-ary Relations

We begin with the basic definition on which the theory of relational databases rests.

DEFINITION 1 Let $A_1, A_2, \ldots, A_n$ be sets. An *n-ary relation* on these sets is a subset of $A_1 \times A_2 \times \cdots \times A_n$. The sets $A_1, A_2, \ldots, A_n$ are called the *domains* of the relation, and n is called its *degree*.

EXAMPLE 1 Let R be the relation on $\mathbf{N} \times \mathbf{N} \times \mathbf{N}$ consisting of triples (a, b, c), where a, b, and c are integers with $a < b < c$. Then $(1, 2, 3) \in R$, but $(2, 4, 3) \notin R$. The degree of this relation is 3. Its domains are all equal to the set of natural numbers. ◀

EXAMPLE 2 Let R be the relation on $\mathbf{Z} \times \mathbf{Z} \times \mathbf{Z}$ consisting of all triples of integers (a, b, c) in which a, b, and c form an arithmetic progression. That is, $(a, b, c) \in R$ if and only if there is an integer k such that $b = a + k$ and $c = a + 2k$, or equivalently, such that $b - a = k$ and $c - b = k$. Note that $(1, 3, 5) \in R$ because $3 = 1 + 2$ and $5 = 1 + 2 \cdot 2$, but $(2, 5, 9) \notin R$ because $5 - 2 = 3$ while $9 - 5 = 4$. This relation has degree 3 and its domains are all equal to the set of integers. ◀

EXAMPLE 3 Let R be the relation on $\mathbf{Z} \times \mathbf{Z} \times \mathbf{Z}^+$ consisting of triples (a, b, m), where a, b, and m are integers with $m \geq 1$ and $a \equiv b \pmod{m}$. Then $(8, 2, 3)$, $(-1, 9, 5)$, and $(14, 0, 7)$ all belong to R, but $(7, 2, 3)$, $(-2, -8, 5)$, and $(11, 0, 6)$ do not belong to R because $8 \equiv 2 \pmod 3$, $-1 \equiv 9 \pmod 5$, and $14 \equiv 0 \pmod 7$, but $7 \not\equiv 2 \pmod 3$, $-2 \not\equiv -8 \pmod 5$, and $11 \not\equiv 0 \pmod 6$. This relation has degree 3 and its first two domains are the set of all integers and its third domain is the set of positive integers. ◀

EXAMPLE 4 Let R be the relation consisting of 5-tuples (A, N, S, D, T) representing airplane flights, where A is the airline, N is the flight number, S is the starting point, D is the destination, and T is the departure time. For instance, if Nadir Express Airlines has flight 963 from Newark to Bangor at 15:00, then (Nadir, 963, Newark, Bangor, 15:00) belongs to R. The degree of this relation is

5, and its domains are the set of all airlines, the set of flight numbers, the set of cities, the set of cities (again), and the set of times. ◀

Databases and Relations

Links

The time required to manipulate information in a database depends on how this information is stored. The operations of adding and deleting records, updating records, searching for records, and combining records from overlapping databases are performed millions of times each day in a large database. Because of the importance of these operations, various methods for representing databases have been developed. We will discuss one of these methods, called the **relational data model,** based on the concept of a relation.

A database consists of **records,** which are n-tuples, made up of **fields.** The fields are the entries of the n-tuples. For instance, a database of student records may be made up of fields containing the name, student number, major, and grade point average of the student. The relational data model represents a database of records as an n-ary relation. Thus, student records are represented as 4-tuples of the form (*STUDENT NAME, ID NUMBER, MAJOR, GPA*). A sample database of six such records is

(Ackermann, 231455, Computer Science, 3.88)
(Adams, 888323, Physics, 3.45)
(Chou, 102147, Computer Science, 3.49)
(Goodfriend, 453876, Mathematics, 3.45)
(Rao, 678543, Mathematics, 3.90)
(Stevens, 786576, Psychology, 2.99).

Relations used to represent databases are also called **tables,** because these relations are often displayed as tables. Each column of the table corresponds to an *attribute* of the database. For instance, the same database of students is displayed in Table 1. The attributes of this database are Student Name, ID Number, Major, and GPA.

A domain of an n-ary relation is called a **primary key** when the value of the n-tuple from this domain determines the n-tuple. That is, a domain is a primary key when no two n-tuples in the relation have the same value from this domain.

Records are often added to or deleted from databases. Because of this, the property that a domain is a primary key is time-dependent. Consequently, a primary key should be chosen that remains one whenever the database is changed. The current collection of n-tuples in a relation is called the **extension** of the relation. The more permanent part of a database, including the name and attributes of the database, is called its **intension.** When selecting a primary key, the goal should be to select a key that can serve as a primary key for all possible extensions of the database. To do this, it is necessary to examine the intension of the database to understand the set of possible n-tuples that can occur in an extension.

TABLE 1 Students.

Student_name	*ID_number*	*Major*	*GPA*
Ackermann	231455	Computer Science	3.88
Adams	888323	Physics	3.45
Chou	102147	Computer Science	3.49
Goodfriend	453876	Mathematics	3.45
Rao	678543	Mathematics	3.90
Stevens	786576	Psychology	2.99

EXAMPLE 5 Which domains are primary keys for the n-ary relation displayed in Table 1, assuming that no n-tuples will be added in the future?

Solution: Because there is only one 4-tuple in this table for each student name, the domain of student names is a primary key. Similarly, the ID numbers in this table are unique, so the domain of ID numbers is also a primary key. However, the domain of major fields of study is not a primary key, because more than one 4-tuple contains the same major field of study. The domain of grade point averages is also not a primary key, because there are two 4-tuples containing the same GPA. ◀

Combinations of domains can also uniquely identify n-tuples in an n-ary relation. When the values of a set of domains determine an n-tuple in a relation, the Cartesian product of these domains is called a **composite key.**

EXAMPLE 6 Is the Cartesian product of the domain of major fields of study and the domain of GPAs a composite key for the n-ary relation from Table 1, assuming that no n-tuples are ever added?

Solution: Because no two 4-tuples from this table have both the same major and the same GPA, this Cartesian product is a composite key. ◀

Because primary and composite keys are used to identify records uniquely in a database, it is important that keys remain valid when new records are added to the database. Hence, checks should be made to ensure that every new record has values that are different in the appropriate field, or fields, from all other records in this table. For instance, it makes sense to use the student identification number as a key for student records if no two students ever have the same student identification number. A university should not use the name field as a key, because two students may have the same name (such as John Smith).

Operations on n-ary Relations

There are a variety of operations on n-ary relations that can be used to form new n-ary relations. Applied together, these operations can answer queries on databases that ask for all tuples that satisfy certain conditions.

The most basic operation on an n-ary relation is determining all n-tuples in the n-ary relation that satisfy certain conditions. For example, we may want to find all the records of all computer science majors in a database of student records. We may want to find all students who have a grade point average above 3.5 in this same database. We may want to find the records of all computer science majors who have a grade point average above 3.5 in this database. To perform such tasks we use the selection operator.

DEFINITION 2 Let R be an n-ary relation and C a condition that elements in R may satisfy. Then the *selection operator* s_C maps the n-ary relation R to the n-ary relation of all n-tuples from R that satisfy the condition C.

EXAMPLE 7 To find the records of computer science majors in the n-ary relation R shown in Table 1, we use the operator s_{C_1}, where C_1 is the condition Major = "Computer Science." The result is the two

4-tuples (Ackermann, 231455, Computer Science, 3.88) and (Chou, 102147, Computer Science, 3.49). Similarly, to find the records of students who have a grade point average above 3.5 in this database, we use the operator s_{C_2}, where C_2 is the condition GPA > 3.5. The result is the two 4-tuples (Ackermann, 231455, Computer Science, 3.88) and (Rao, 678543, Mathematics, 3.90). Finally, to find the records of computer science majors who have a GPA above 3.5, we use the operator s_{C_3}, where C_3 is the condition (Major = "Computer Science" $\wedge$ GPA > 3.5). The result consists of the single 4-tuple (Ackermann, 231455, Computer Science, 3.88). ◀

Projections are used to form new n-ary relations by deleting the same fields in every record of the relation.

DEFINITION 3 The *projection* $P_{i_1 i_2, \ldots, i_m}$ where $i_1 < i_2 < \cdots < i_m$, maps the n-tuple $(a_1, a_2, \ldots, a_n)$ to the m-tuple $(a_{i_1}, a_{i_2}, \ldots, a_{i_m})$, where $m \leq n$.

In other words, the projection $P_{i_1, i_2, \ldots, i_m}$ deletes $n - m$ of the components of an n-tuple, leaving the i_1th, i_2th, ..., and i_mth components.

EXAMPLE 8 What results when the projection $P_{1,3}$ is applied to the 4-tuples $(2, 3, 0, 4)$, (Jane Doe, 234111001, Geography, 3.14), and (a_1, a_2, a_3, a_4)?

Solution: The projection $P_{1,3}$ sends these 4-tuples to $(2, 0)$, (Jane Doe, Geography), and (a_1, a_3), respectively. ◀

Example 9 illustrates how new relations are produced using projections.

EXAMPLE 9 What relation results when the projection $P_{1,4}$ is applied to the relation in Table 1?

Solution: When the projection $P_{1,4}$ is used, the second and third columns of the table are deleted, and pairs representing student names and grade point averages are obtained. Table 2 displays the results of this projection. ◀

Fewer rows may result when a projection is applied to the table for a relation. This happens when some of the n-tuples in the relation have identical values in each of the m components of the projection, and only disagree in components deleted by the projection. For instance, consider the following example.

EXAMPLE 10 What is the table obtained when the projection $P_{1,2}$ is applied to the relation in Table 3?

Solution: Table 4 displays the relation obtained when $P_{1,2}$ is applied to Table 3. Note that there are fewer rows after this projection is applied. ◀

The **join** operation is used to combine two tables into one when these tables share some identical fields. For instance, a table containing fields for airline, flight number, and gate, and another table containing fields for flight number, gate, and departure time can be combined into a table containing fields for airline, flight number, gate, and departure time.

TABLE 2 GPAs.

Student_name	*GPA*
Ackermann	3.88
Adams	3.45
Chou	3.49
Goodfriend	3.45
Rao	3.90
Stevens	2.99

TABLE 3 Enrollments.

Student	*Major*	*Course*
Glauser	Biology	BI 290
Glauser	Biology	MS 475
Glauser	Biology	PY 410
Marcus	Mathematics	MS 511
Marcus	Mathematics	MS 603
Marcus	Mathematics	CS 322
Miller	Computer Science	MS 575
Miller	Computer Science	CS 455

TABLE 4 Majors.

Student	*Major*
Glauser	Biology
Marcus	Mathematics
Miller	Computer Science

DEFINITION 4 Let R be a relation of degree m and S a relation of degree n. The *join* $J_p(R, S)$, where $p \leq m$ and $p \leq n$, is a relation of degree $m + n - p$ that consists of all $(m + n - p)$-tuples $(a_1, a_2, \ldots, a_{m-p}, c_1, c_2, \ldots, c_p, b_1, b_2, \ldots, b_{n-p})$, where the m-tuple $(a_1, a_2, \ldots, a_{m-p}, c_1, c_2, \ldots, c_p)$ belongs to R and the n-tuple $(c_1, c_2, \ldots, c_p, b_1, b_2, \ldots, b_{n-p})$ belongs to S.

In other words, the join operator J_p produces a new relation from two relations by combining all m-tuples of the first relation with all n-tuples of the second relation, where the last p components of the m-tuples agree with the first p components of the n-tuples.

EXAMPLE 11 What relation results when the join operator J_2 is used to combine the relation displayed in Tables 5 and 6?

Solution: The join J_2 produces the relation shown in Table 7. ◀

There are other operators besides projections and joins that produce new relations from existing relations. A description of these operations can be found in books on database theory.

TABLE 5 Teaching_assignments.

Professor	*Department*	*Course_number*
Cruz	Zoology	335
Cruz	Zoology	412
Farber	Psychology	501
Farber	Psychology	617
Grammer	Physics	544
Grammer	Physics	551
Rosen	Computer Science	518
Rosen	Mathematics	575

TABLE 6 Class_schedule.

Department	*Course_number*	*Room*	*Time*
Computer Science	518	N521	2:00 P.M.
Mathematics	575	N502	3:00 P.M.
Mathematics	611	N521	4:00 P.M.
Physics	544	B505	4:00 P.M.
Psychology	501	A100	3:00 P.M.
Psychology	617	A110	11:00 A.M.
Zoology	335	A100	9:00 A.M.
Zoology	412	A100	8:00 A.M.

TABLE 7 Teaching_schedule.

Professor	*Department*	*Course_number*	*Room*	*Time*
Cruz	Zoology	335	A100	9:00 A.M.
Cruz	Zoology	412	A100	8:00 A.M.
Farber	Psychology	501	A100	3:00 P.M.
Farber	Psychology	617	A110	11:00 A.M.
Grammer	Physics	544	B505	4:00 P.M.
Rosen	Computer Science	518	N521	2:00 P.M.
Rosen	Mathematics	575	N502	3:00 P.M.

SQL

Links

The database query language SQL (short for structured query language) can be used to carry out the operations we have described in this section. Example 12 illustrates how SQL commands are related to operations on n-ary relations.

EXAMPLE 12 We will illustrate how SQL is used to express queries by showing how SQL can be employed to make a query about airline flights using Table 8. The SQL statement

```
SELECT Departure_time
FROM Flights
WHERE Destination='Detroit'
```

is used to find the projection P_5 (on the Departure_time attribute) of the selection of 5-tuples in the Flights database that satisfy the condition: Destination = 'Detroit'. The output would be a list containing the times of flights that have Detroit as their destination, namely, 08:10, 08:47, and 09:44. SQL uses the FROM clause to identify the n-ary relation the query is applied to, the WHERE clause to specify the condition of the selection operation, and the SELECT clause to specify the projection operation that is to be applied. (*Beware:* SQL uses SELECT to represent a projection, rather than a selection operation. This is an unfortunate example of conflicting terminology.) ◀

Example 13 shows how SQL queries can be made involving more than one table.

TABLE 8 Flights.

Airline	*Flight_number*	*Gate*	*Destination*	*Departure_time*
Nadir	122	34	Detroit	08:10
Acme	221	22	Denver	08:17
Acme	122	33	Anchorage	08:22
Acme	323	34	Honolulu	08:30
Nadir	199	13	Detroit	08:47
Acme	222	22	Denver	09:10
Nadir	322	34	Detroit	09:44

EXAMPLE 13 The SQL statement

```
SELECT Professor, Time
FROM Teaching_assignments, Class_schedule
WHERE Department='Mathematics'
```

is used to find the projection $P_{1,5}$ of the 5-tuples in the database (shown in Table 7), which is the join J_2 of the Teaching_assignments and Class_schedule databases in Tables 5 and 6, respectively, which satisfy the condition: Department = Mathematics. The output would consist of the single 2-tuple (Rosen, 3:00 P.M.). The SQL FROM clause is used here to find the join of two different databases. ◀

We have only touched on the basic concepts of relational databases in this section. More information can be found in [AhUl95].

Exercises

1. List the triples in the relation $\{(a, b, c) \mid a, b, \text{ and } c \text{ are integers with } 0 < a < b < c < 5\}$.
2. Which 4-tuples are in the relation $\{(a, b, c, d) \mid a, b, c, \text{ and } d \text{ are positive integers with } abcd = 6\}$?
3. List the 5-tuples in the relation in Table 8.
4. Assuming that no new n-tuples are added, find all the primary keys for the relations displayed in
   a) Table 3. **b)** Table 5.
   c) Table 6. **d)** Table 8.
5. Assuming that no new n-tuples are added, find a composite key with two fields containing the *Airline* field for the database in Table 8.
6. Assuming that no new n-tuples are added, find a composite key with two fields containing the Professor field for the database in Table 7.
7. The 3-tuples in a 3-ary relation represent the following attributes of a student database: student ID number, name, phone number.
   a) Is student ID number likely to be a primary key?
   b) Is name likely to be a primary key?
   c) Is phone number likely to be a primary key?
8. The 4-tuples in a 4-ary relation represent these attributes of published books: title, ISBN, publication date, number of pages.
   a) What is a likely primary key for this relation?
   b) Under what conditions would (title, publication date) be a composite key?
   c) Under what conditions would (title, number of pages) be a composite key?
9. The 5-tuples in a 5-ary relation represent these attributes of all people in the United States: name, Social Security number, street address, city, state.
   a) Determine a primary key for this relation.
   b) Under what conditions would (name, street address) be a composite key?
   c) Under what conditions would (name, street address, city) be a composite key?
10. What do you obtain when you apply the selection operator s_C, where C is the condition Room = A100, to the database in Table 7?
11. What do you obtain when you apply the selection operator s_C, where C is the condition Destination = Detroit, to the database in Table 8?
12. What do you obtain when you apply the selection operator s_C, where C is the condition (Project = 2) $\wedge$ (Quantity $\geq$ 50), to the database in Table 10 (found later in this exercise set)?
13. What do you obtain when you apply the selection operator s_C, where C is the condition (Airline = Nadir) $\vee$ (Destination = Denver), to the database in Table 8?
14. What do you obtain when you apply the projection $P_{2,3,5}$ to the 5-tuple (a, b, c, d, e)?
15. Which projection mapping is used to delete the first, second, and fourth components of a 6-tuple?
16. Display the table produced by applying the projection $P_{1,2,4}$ to Table 8.
17. Display the table produced by applying the projection $P_{1,4}$ to Table 8.
18. How many components are there in the n-tuples in the table obtained by applying the join operator J_3 to two tables with 5-tuples and 8-tuples, respectively?
19. Construct the table obtained by applying the join operator J_2 to the relations in Tables 9 and 10.
20. Show that if C_1 and C_2 are conditions that elements of the n-ary relation R may satisfy, then $s_{C_1 \wedge C_2}(R) = s_{C_1}(s_{C_2}(R))$.
21. Show that if C_1 and C_2 are conditions that elements of the n-ary relation R may satisfy, then $s_{C_1}(s_{C_2}(R)) = s_{C_2}(s_{C_1}(R))$.
22. Show that if C is a condition that elements of the n-ary relations R and S may satisfy, then $s_C(R \cup S) = s_C(R) \cup s_C(S)$.
23. Show that if C is a condition that elements of the n-ary relations R and S may satisfy, then $s_C(R \cap S) = s_C(R) \cap s_C(S)$.

TABLE 9 Part_needs.

Supplier	*Part_number*	*Project*
23	1092	1
23	1101	3
23	9048	4
31	4975	3
31	3477	2
32	6984	4
32	9191	2
33	1001	1

TABLE 10 Parts_inventory.

Part_number	*Project*	*Quantity*	*Color_code*
1001	1	14	8
1092	1	2	2
1101	3	1	1
3477	2	25	2
4975	3	6	2
6984	4	10	1
9048	4	12	2
9191	2	80	4

24. Show that if C is a condition that elements of the n-ary relations R and S may satisfy, then $s_C(R - S) = s_C(R) - s_C(S)$.

25. Show that if R and S are both n-ary relations, then $P_{i_1,i_2,\ldots,i_m}(R \cup S) = P_{i_1,i_2,\ldots,i_m}(R) \cup P_{i_1,i_2,\ldots,i_m}(S)$.

26. Give an example to show that if R and S are both n-ary relations, then $P_{i_1,i_2,\ldots,i_m}(R \cap S)$ may be different from $P_{i_1,i_2,\ldots,i_m}(R) \cap P_{i_1,i_2,\ldots,i_m}(S)$.

27. Give an example to show that if R and S are both n-ary relations, then $P_{i_1,i_2,\ldots,i_m}(R - S)$ may be different from $P_{i_1,i_2,\ldots,i_m}(R) - P_{i_1,i_2,\ldots,i_m}(S)$.

28. a) What are the operations that correspond to the query expressed using this SQL statement?

```
SELECT Supplier
FROM Part_needs
WHERE 1000 ≤ Part_number ≤ 5000
```

b) What is the output of this query given the database in Table 9 as input?

29. a) What are the operations that correspond to the query expressed using this SQL statement?

```
SELECT Supplier, Project
FROM Part_needs, Parts_inventory
WHERE Quantity ≤ 10
```

b) What is the output of this query given the databases in Tables 9 and 10 as input?

30. Determine whether there is a primary key for the relation in Example 2.

31. Determine whether there is a primary key for the relation in Example 3.

32. Show that an n-ary relation with a primary key can be thought of as the graph of a function that maps values of the primary key to $(n - 1)$-tuples formed from values of the other domains.

8.3 Representing Relations

Introduction

In this section, and in the remainder of this chapter, all relations we study will be binary relations. Because of this, in this section and in the rest of this chapter, the word relation will always refer to a binary relation. There are many ways to represent a relation between finite sets. As we have seen in Section 8.1, one way is to list its ordered pairs. Another way to represent a relation is to use a table, as we did in Example 3 in Section 8.1. In this section we will discuss two alternative methods for representing relations. One method uses zero–one matrices. The other method uses pictorial representations called directed graphs, which we will discuss later in this section.

Generally, matrices are appropriate for the representation of relations in computer programs. On the other hand, people often find the representation of relations using directed graphs useful for understanding the properties of these relations.

Representing Relations Using Matrices

A relation between finite sets can be represented using a zero–one matrix. Suppose that R is a relation from $A = \{a_1, a_2, \ldots, a_m\}$ to $B = \{b_1, b_2, \ldots, b_n\}$. (Here the elements of the sets A and B have been listed in a particular, but arbitrary, order. Furthermore, when $A = B$ we use the same ordering for A and B.) The relation R can be represented by the matrix $\mathbf{M}_R = [m_{ij}]$, where

$$m_{ij} = \begin{cases} 1 \text{ if } (a_i, b_j) \in R, \\ 0 \text{ if } (a_i, b_j) \notin R. \end{cases}$$

In other words, the zero–one matrix representing R has a 1 as its (i, j) entry when a_i is related to b_j, and a 0 in this position if a_i is not related to b_j. (Such a representation depends on the orderings used for A and B.)

The use of matrices to represent relations is illustrated in Examples 1–6.

EXAMPLE 1 Suppose that $A = \{1, 2, 3\}$ and $B = \{1, 2\}$. Let R be the relation from A to B containing (a, b) if $a \in A$, $b \in B$, and $a > b$. What is the matrix representing R if $a_1 = 1$, $a_2 = 2$, and $a_3 = 3$, and $b_1 = 1$ and $b_2 = 2$?

Solution: Because $R = \{(2, 1), (3, 1), (3, 2)\}$, the matrix for R is

$$\mathbf{M}_R = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

The 1s in $\mathbf{M}_R$ show that the pairs (2, 1), (3, 1), and (3, 2) belong to R. The 0s show that no other pairs belong to R. ◀

EXAMPLE 2 Let $A = \{a_1, a_2, a_3\}$ and $B = \{b_1, b_2, b_3, b_4, b_5\}$. Which ordered pairs are in the relation R represented by the matrix

$$\mathbf{M}_R = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix}?$$

Solution: Because R consists of those ordered pairs (a_i, b_j) with $m_{ij} = 1$, it follows that

$$R = \{(a_1, b_2), (a_2, b_1), (a_2, b_3), (a_2, b_4), (a_3, b_1), (a_3, b_3), (a_3, b_5)\}.$$ ◀

The matrix of a relation on a set, which is a square matrix, can be used to determine whether the relation has certain properties. Recall that a relation R on A is reflexive if $(a, a) \in R$ whenever $a \in A$. Thus, R is reflexive if and only if $(a_i, a_i) \in R$ for $i = 1, 2, \ldots, n$. Hence, R is reflexive if and only if $m_{ii} = 1$, for $i = 1, 2, \ldots, n$. In other words, R is reflexive if all the elements on the main diagonal of $\mathbf{M}_R$ are equal to 1, as shown in Figure 1.

$$\begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & 1 & & & & \\ & & & \cdot & & & \\ & & & & \cdot & & \\ & & & & & \cdot & \\ & & & & & & 1 \\ & & & & & & & 1 \end{bmatrix}$$

FIGURE 1 The Zero–One Matrix for a Reflexive Relation.

The relation R is symmetric if $(a, b) \in R$ implies that $(b, a) \in R$. Consequently, the relation R on the set $A = \{a_1, a_2, \ldots, a_n\}$ is symmetric if and only if $(a_j, a_i) \in R$ whenever $(a_i, a_j) \in R$. In terms of the entries of $\mathbf{M}_R$, R is symmetric if and only if $m_{ji} = 1$ whenever $m_{ij} = 1$. This also means $m_{ji} = 0$ whenever $m_{ij} = 0$. Consequently, R is symmetric if and only if $m_{ij} = m_{ji}$, for all pairs of integers i and j with $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$. Recalling the definition of the transpose of a matrix from Section 3.8, we see that R is symmetric if and only if

$$\mathbf{M}_R = (\mathbf{M}_R)^t,$$

(a) Symmetric (b) Antisymmetric

FIGURE 2 The Zero–One Matrices for Symmetric and Antisymmetric Relations.

that is, if $\mathbf{M}_R$ is a symmetric matrix. The form of the matrix for a symmetric relation is illustrated in Figure 2(a).

The relation R is antisymmetric if and only if $(a, b) \in R$ and $(b, a) \in R$ imply that $a = b$. Consequently, the matrix of an antisymmetric relation has the property that if $m_{ij} = 1$ with $i \neq j$, then $m_{ji} = 0$. Or, in other words, either $m_{ij} = 0$ or $m_{ji} = 0$ when $i \neq j$. The form of the matrix for an antisymmetric relation is illustrated in Figure 2(b).

EXAMPLE 3 Suppose that the relation R on a set is represented by the matrix

$$\mathbf{M}_R = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Is R reflexive, symmetric, and/or antisymmetric?

Solution: Because all the diagonal elements of this matrix are equal to 1, R is reflexive. Moreover, because $\mathbf{M}_R$ is symmetric, it follows that R is symmetric. It is also easy to see that R is not antisymmetric. ◀

The Boolean operations join and meet (discussed in Section 3.8) can be used to find the matrices representing the union and the intersection of two relations. Suppose that R_1 and R_2 are relations on a set A represented by the matrices $\mathbf{M}_{R_1}$ and $\mathbf{M}_{R_2}$, respectively. The matrix representing the union of these relations has a 1 in the positions where either $\mathbf{M}_{R_1}$ or $\mathbf{M}_{R_2}$ has a 1. The matrix representing the intersection of these relations has a 1 in the positions where both $\mathbf{M}_{R_1}$ and $\mathbf{M}_{R_2}$ have a 1. Thus, the matrices representing the union and intersection of these relations are

$$\mathbf{M}_{R_1 \cup R_2} = \mathbf{M}_{R_1} \vee \mathbf{M}_{R_2} \qquad \text{and} \qquad \mathbf{M}_{R_1 \cap R_2} = \mathbf{M}_{R_1} \wedge \mathbf{M}_{R_2}.$$

EXAMPLE 4 Suppose that the relations R_1 and R_2 on a set A are represented by the matrices

$$\mathbf{M}_{R_1} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \qquad \text{and} \qquad \mathbf{M}_{R_2} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

What are the matrices representing $R_1 \cup R_2$ and $R_1 \cap R_2$?

Solution: The matrices of these relations are

$$\mathbf{M}_{R_1 \cup R_2} = \mathbf{M}_{R_1} \vee \mathbf{M}_{R_2} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix},$$

$$\mathbf{M}_{R_1 \cap R_2} = \mathbf{M}_{R_1} \wedge \mathbf{M}_{R_2} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

◀

We now turn our attention to determining the matrix for the composite of relations. This matrix can be found using the Boolean product of the matrices (discussed in Section 3.8) for these relations. In particular, suppose that R is a relation from A to B and S is a relation from B to C. Suppose that A, B, and C have m, n, and p elements, respectively. Let the zero–one matrices for $S \circ R$, R, and S be $\mathbf{M}_{S \circ R} = [t_{ij}]$, $\mathbf{M}_R = [r_{ij}]$, and $\mathbf{M}_S = [s_{ij}]$, respectively (these matrices have sizes $m \times p$, $m \times n$, and $n \times p$, respectively). The ordered pair (a_i, c_j) belongs to $S \circ R$ if and only if there is an element b_k such that (a_i, b_k) belongs to R and (b_k, c_j) belongs to S. It follows that $t_{ij} = 1$ if and only if $r_{ik} = s_{kj} = 1$ for some k. From the definition of the Boolean product, this means that

$$\mathbf{M}_{S \circ R} = \mathbf{M}_R \odot \mathbf{M}_S.$$

EXAMPLE 5 Find the matrix representing the relations $S \circ R$, where the matrices representing R and S are

$$\mathbf{M}_R = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{M}_S = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

Solution: The matrix for $S \circ R$ is

$$\mathbf{M}_{S \circ R} = \mathbf{M}_R \odot \mathbf{M}_S = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

◀

The matrix representing the composite of two relations can be used to find the matrix for $\mathbf{M}_{R^n}$. In particular,

$$\mathbf{M}_{R^n} = \mathbf{M}_R^{[n]},$$

from the definition of Boolean powers. Exercise 35 at the end of this section asks for a proof of this formula.

EXAMPLE 6 Find the matrix representing the relation R^2, where the matrix representing R is

$$\mathbf{M}_R = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Solution: The matrix for R^2 is

$$\mathbf{M}_{R^2} = \mathbf{M}_R^{[2]} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

◀

Representing Relations Using Digraphs

We have shown that a relation can be represented by listing all of its ordered pairs or by using a zero–one matrix. There is another important way of representing a relation using a pictorial representation. Each element of the set is represented by a point, and each ordered pair is represented using an arc with its direction indicated by an arrow. We use such pictorial representations when we think of relations on a finite set as **directed graphs,** or **digraphs.**

DEFINITION 1 A *directed graph,* or *digraph,* consists of a set V of *vertices* (or *nodes*) together with a set E of ordered pairs of elements of V called *edges* (or *arcs*). The vertex a is called the *initial vertex* of the edge (a, b), and the vertex b is called the *terminal vertex* of this edge.

An edge of the form (a, a) is represented using an arc from the vertex a back to itself. Such an edge is called a **loop.**

EXAMPLE 7 The directed graph with vertices a, b, c, and d, and edges (a, b), (a, d), (b, b), (b, d), (c, a), (c, b), and (d, b) is displayed in Figure 3. ◀

FIGURE 3
A Directed Graph.

The relation R on a set A is represented by the directed graph that has the elements of A as its vertices and the ordered pairs (a, b), where $(a, b) \in R$, as edges. This assignment sets up a one-to-one correspondence between the relations on a set A and the directed graphs with A as their set of vertices. Thus, every statement about relations corresponds to a statement about directed graphs, and vice versa. Directed graphs give a visual display of information about relations. As such, they are often used to study relations and their properties. (Note that relations from a set A to a set B can be represented by a directed graph where there is a vertex for each element of A and a vertex for each element of B, as shown in Section 8.1. However, when $A = B$, such representation provides much less insight than the digraph representations described here.) The use of directed graphs to represent relations on a set is illustrated in Examples 8–10.

EXAMPLE 8 The directed graph of the relation

$$R = \{(1, 1), (1, 3), (2, 1), (2, 3), (2, 4), (3, 1), (3, 2), (4, 1)\}$$

on the set $\{1, 2, 3, 4\}$ is shown in Figure 4. ◀

EXAMPLE 9 What are the ordered pairs in the relation R represented by the directed graph shown in Figure 5?

Solution: The ordered pairs (x, y) in the relation are

$$R = \{(1, 3), (1, 4), (2, 1), (2, 2), (2, 3), (3, 1), (3, 3), (4, 1), (4, 3)\}.$$

Each of these pairs corresponds to an edge of the directed graph, with $(2, 2)$ and $(3, 3)$ corresponding to loops. ◀

The directed graph representing a relation can be used to determine whether the relation has various properties. For instance, a relation is reflexive if and only if there is a loop at every vertex of the directed graph, so that every ordered pair of the form (x, x) occurs in the relation. A relation is symmetric if and only if for every edge between distinct vertices in its digraph there is an edge in the opposite direction, so that (y, x) is in the relation whenever (x, y) is in the

FIGURE 4 The Directed Graph of the Relation *R*.

FIGURE 5 The Directed Graph of the Relation *R*.

(a) Directed graph of R (b) Directed graph of S

FIGURE 6 The Directed Graphs of the Relations *R* and *S*.

relation. Similarly, a relation is antisymmetric if and only if there are never two edges in opposite directions between distinct vertices. Finally, a relation is transitive if and only if whenever there is an edge from a vertex x to a vertex y and an edge from a vertex y to a vertex z, there is an edge from x to z (completing a triangle where each side is a directed edge with the correct direction).

Remark: Note that a symmetric relation can be represented by an undirected graph, which is a graph where edges do not have directions. We will study undirected graphs in Chapter 9.

EXAMPLE 10 Determine whether the relations for the directed graphs shown in Figure 6 are reflexive, symmetric, antisymmetric, and/or transitive.

Solution: Because there are loops at every vertex of the directed graph of R, it is reflexive. R is neither symmetric nor antisymmetric because there is an edge from a to b but not one from b to a, but there are edges in both directions connecting b and c. Finally, R is not transitive because there is an edge from a to b and an edge from b to c, but no edge from a to c.

Because loops are not present at all the vertices of the directed graph of S, this relation is not reflexive. It is symmetric and not antisymmetric, because every edge between distinct vertices is accompanied by an edge in the opposite direction. It is also not hard to see from the directed graph that S is not transitive, because (c, a) and (a, b) belong to S, but (c, b) does not belong to S. ◀

Exercises

1. Represent each of these relations on $\{1, 2, 3\}$ with a matrix (with the elements of this set listed in increasing order).
   a) $\{(1, 1), (1, 2), (1, 3)\}$
   b) $\{(1, 2), (2, 1), (2, 2), (3, 3)\}$
   c) $\{(1, 1), (1, 2), (1, 3), (2, 2), (2, 3), (3, 3)\}$
   d) $\{(1, 3), (3, 1)\}$
2. Represent each of these relations on $\{1, 2, 3, 4\}$ with a matrix (with the elements of this set listed in increasing order).
   a) $\{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}$
   b) $\{(1, 1), (1, 4), (2, 2), (3, 3), (4, 1)\}$
   c) $\{(1, 2), (1, 3), (1, 4), (2, 1), (2, 3), (2, 4), (3, 1), (3, 2), (3, 4), (4, 1), (4, 2), (4, 3)\}$
   d) $\{(2, 4), (3, 1), (3, 2), (3, 4)\}$
3. List the ordered pairs in the relations on $\{1, 2, 3\}$ corresponding to these matrices (where the rows and columns correspond to the integers listed in increasing order).
   a) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ **b)** $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$
   c) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$
4. List the ordered pairs in the relations on $\{1, 2, 3, 4\}$ corresponding to these matrices (where the rows and

columns correspond to the integers listed in increasing order).

a) $\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$

c) $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$

5. How can the matrix representing a relation R on a set A be used to determine whether the relation is irreflexive?
6. How can the matrix representing a relation R on a set A be used to determine whether the relation is asymmetric?
7. Determine whether the relations represented by the matrices in Exercise 3 are reflexive, irreflexive, symmetric, antisymmetric, and/or transitive.
8. Determine whether the relations represented by the matrices in Exercise 4 are reflexive, irreflexive, symmetric, antisymmetric, and/or transitive.
9. How many nonzero entries does the matrix representing the relation R on $A = \{1, 2, 3, \ldots, 100\}$ consisting of the first 100 positive integers have if R is
   a) $\{(a, b) \mid a > b\}$? b) $\{(a, b) \mid a \neq b\}$?
   c) $\{(a, b) \mid a = b + 1\}$? d) $\{(a, b) \mid a = 1\}$?
   e) $\{(a, b) \mid ab = 1\}$?
10. How many nonzero entries does the matrix representing the relation R on $A = \{1, 2, 3, \ldots, 1000\}$ consisting of the first 1000 positive integers have if R is
   a) $\{(a, b) \mid a \leq b\}$?
   b) $\{(a, b) \mid a = b \pm 1\}$?
   c) $\{(a, b) \mid a + b = 1000\}$?
   d) $\{(a, b) \mid a + b \leq 1001\}$?
   e) $\{(a, b) \mid a \neq 0\}$?
11. How can the matrix for $\overline{R}$, the complement of the relation R, be found from the matrix representing R, when R is a relation on a finite set A?
12. How can the matrix for R^{-1}, the inverse of the relation R, be found from the matrix representing R, when R is a relation on a finite set A?
13. Let R be the relation represented by the matrix

$$\mathbf{M}_R = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

Find the matrix representing
a) R^{-1}. b) $\overline{R}$. c) R^2.

14. Let R_1 and R_2 be relations on a set A represented by the matrices

$$\mathbf{M}_{R_1} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{M}_{R_2} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Find the matrices that represent
a) $R_1 \cup R_2$. b) $R_1 \cap R_2$. c) $R_2 \circ R_1$.
d) $R_1 \circ R_1$. e) $R_1 \oplus R_2$.

15. Let R be the relation represented by the matrix

$$\mathbf{M}_R = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

Find the matrices that represent
a) R^2. b) R^3. c) R^4.

16. Let R be a relation on a set A with n elements. If there are k nonzero entries in $\mathbf{M}_R$, the matrix representing R, how many nonzero entries are there in $\mathbf{M}_{R^{-1}}$, the matrix representing R^{-1}, the inverse of R?
17. Let R be a relation on a set A with n elements. If there are k nonzero entries in $\mathbf{M}_R$, the matrix representing R, how many nonzero entries are there in $\mathbf{M}_{\overline{R}}$, the matrix representing $\overline{R}$, the complement of R?
18. Draw the directed graphs representing each of the relations from Exercise 1.
19. Draw the directed graphs representing each of the relations from Exercise 2.
20. Draw the directed graph representing each of the relations from Exercise 3.
21. Draw the directed graph representing each of the relations from Exercise 4.
22. Draw the directed graph that represents the relation $\{(a, a), (a, b), (b, c), (c, b), (c, d), (d, a), (d, b)\}$.

In Exercises 23–28 list the ordered pairs in the relations represented by the directed graphs.

23.

24.

25.

26.

27.

28.

29. How can the directed graph of a relation R on a finite set A be used to determine whether a relation is asymmetric?

30. How can the directed graph of a relation R on a finite set A be used to determine whether a relation is irreflexive?

31. Determine whether the relations represented by the directed graphs shown in Exercises 23–25 are reflexive, irreflexive, symmetric, antisymmetric, and/or transitive.

32. Determine whether the relations represented by the directed graphs shown in Exercises 26–28 are reflexive, irreflexive, symmetric, antisymmetric, asymmetric, and/or transitive.

33. Let R be a relation on a set A. Explain how to use the directed graph representing R to obtain the directed graph representing the inverse relation R^{-1}.

34. Let R be a relation on a set A. Explain how to use the directed graph representing R to obtain the directed graph representing the complementary relation $\overline{R}$.

35. Show that if $\mathbf{M}_R$ is the matrix representing the relation R, then $\mathbf{M}_R^{[n]}$ is the matrix representing the relation R^n.

36. Given the directed graphs representing two relations, how can the directed graph of the union, intersection, symmetric difference, difference, and composition of these relations be found?

8.4 Closures of Relations

Introduction

A computer network has data centers in Boston, Chicago, Denver, Detroit, New York, and San Diego. There are direct, one-way telephone lines from Boston to Chicago, from Boston to Detroit, from Chicago to Detroit, from Detroit to Denver, and from New York to San Diego. Let R be the relation containing (a, b) if there is a telephone line from the data center in a to that in b. How can we determine if there is some (possibly indirect) link composed of one or more telephone lines from one center to another? Because not all links are direct, such as the link from Boston to Denver that goes through Detroit, R cannot be used directly to answer this. In the language of relations, R is not transitive, so it does not contain all the pairs that can be linked. As we will show in this section, we can find all pairs of data centers that have a link by constructing a transitive relation S containing R such that S is a subset of every transitive relation containing R. Here, S is the smallest transitive relation that contains R. This relation is called the **transitive closure** of R.

In general, let R be a relation on a set A. R may or may not have some property **P**, such as reflexivity, symmetry, or transitivity. If there is a relation S with property **P** containing R such that S is a subset of every relation with property **P** containing R, then S is called the **closure** of R with respect to **P**. (Note that the closure of a relation with respect to a property may not exist; see Exercises 15 and 35 at the end of this section.) We will show how reflexive, symmetric, and transitive closures of relations can be found.

Closures

The relation $R = \{(1, 1), (1, 2), (2, 1), (3, 2)\}$ on the set $A = \{1, 2, 3\}$ is not reflexive. How can we produce a reflexive relation containing R that is as small as possible? This can be done by adding $(2, 2)$ and $(3, 3)$ to R, because these are the only pairs of the form (a, a) that are not in R. Clearly, this new relation contains R. Furthermore, *any* reflexive relation that contains R must also contain $(2, 2)$ and $(3, 3)$. Because this relation contains R, is reflexive, and is contained within every reflexive relation that contains R, it is called the **reflexive closure** of R.

As this example illustrates, given a relation R on a set A, the reflexive closure of R can be formed by adding to R all pairs of the form (a, a) with $a \in A$, not already in R. The

addition of these pairs produces a new relation that is reflexive, contains R, and is contained within any reflexive relation containing R. We see that the reflexive closure of R equals $R \cup \Delta$, where $\Delta = \{(a, a) \mid a \in A\}$ is the **diagonal relation** on A. (The reader should verify this.)

EXAMPLE 1 What is the reflexive closure of the relation $R = \{(a, b) \mid a < b\}$ on the set of integers?

Solution: The reflexive closure of R is

$$R \cup \Delta = \{(a, b) \mid a < b\} \cup \{(a, a) \mid a \in \mathbf{Z}\} = \{(a, b) \mid a \leq b\}.$$ ◀

The relation $\{(1, 1), (1, 2), (2, 2), (2, 3), (3, 1), (3, 2)\}$ on $\{1, 2, 3\}$ is not symmetric. How can we produce a symmetric relation that is as small as possible and contains R? To do this, we need only add $(2, 1)$ and $(1, 3)$, because these are the only pairs of the form (b, a) with $(a, b) \in R$ that are not in R. This new relation is symmetric and contains R. Furthermore, *any* symmetric relation that contains R must contain this new relation, because a symmetric relation that contains R must contain $(2, 1)$ and $(1, 3)$. Consequently, this new relation is called the **symmetric closure** of R.

As this example illustrates, the symmetric closure of a relation R can be constructed by adding all ordered pairs of the form (b, a), where (a, b) is in the relation, that are not already present in R. Adding these pairs produces a relation that is symmetric, that contains R, and that is contained in any symmetric relation that contains R. The symmetric closure of a relation can be constructed by taking the union of a relation with its inverse (defined in the preamble of Exercise 24 in Section 8.1); that is, $R \cup R^{-1}$ is the symmetric closure of R, where $R^{-1} = \{(b, a) \mid (a, b) \in R\}$. The reader should verify this statement.

EXAMPLE 2 What is the symmetric closure of the relation $R = \{(a, b) \mid a > b\}$ on the set of positive integers?

Extra Examples

Solution: The symmetric closure of R is the relation

$$R \cup R^{-1} = \{(a, b) \mid a > b\} \cup \{(b, a) \mid a > b\} = \{(a, b) \mid a \neq b\}.$$

This last equality follows because R contains all ordered pairs of positive integers where the first element is greater than the second element and R^{-1} contains all ordered pairs of positive integers where the first element is less than the second. ◀

Suppose that a relation R is not transitive. How can we produce a transitive relation that contains R such that this new relation is contained within any transitive relation that contains R? Can the transitive closure of a relation R be produced by adding all the pairs of the form (a, c), where (a, b) and (b, c) are already in the relation? Consider the relation $R = \{(1, 3), (1, 4), (2, 1), (3, 2)\}$ on the set $\{1, 2, 3, 4\}$. This relation is not transitive because it does not contain all pairs of the form (a, c) where (a, b) and (b, c) are in R. The pairs of this form not in R are $(1, 2)$, $(2, 3)$, $(2, 4)$, and $(3, 1)$. Adding these pairs does *not* produce a transitive relation, because the resulting relation contains $(3, 1)$ and $(1, 4)$ but does not contain $(3, 4)$. This shows that constructing the transitive closure of a relation is more complicated than constructing either the reflexive or symmetric closure. The rest of this section develops algorithms for constructing transitive closures. As will be shown later in this section, the transitive closure of a relation can be found by adding new ordered pairs that must be present and then repeating this process until no new ordered pairs are needed.

Paths in Directed Graphs

We will see that representing relations by directed graphs helps in the construction of transitive closures. We now introduce some terminology that we will use for this purpose.

A path in a directed graph is obtained by traversing along edges (in the same direction as indicated by the arrow on the edge).

DEFINITION 1 A *path* from a to b in the directed graph G is a sequence of edges (x_0, x_1), (x_1, x_2), $(x_2, x_3), \ldots, (x_{n-1}, x_n)$ in G, where n is a nonnegative integer, and $x_0 = a$ and $x_n = b$, that is, a sequence of edges where the terminal vertex of an edge is the same as the initial vertex in the next edge in the path. This path is denoted by $x_0, x_1, x_2, \ldots, x_{n-1}, x_n$ and has *length* n. We view the empty set of edges as a path from a to a. A path of length $n \geq 1$ that begins and ends at the same vertex is called a *circuit* or *cycle*.

A path in a directed graph can pass through a vertex more than once. Moreover, an edge in a directed graph can occur more than once in a path.

EXAMPLE 3 Which of the following are paths in the directed graph shown in Figure 1: a, b, e, d; a, e, c, d, b; b, a, c, b, a, a, b; d, c; c, b, a; e, b, a, b, a, b, e? What are the lengths of those that are paths? Which of the paths in this list are circuits?

Solution: Because each of (a, b), (b, e), and (e, d) is an edge, a, b, e, d is a path of length three. Because (c, d) is not an edge, a, e, c, d, b is not a path. Also, b, a, c, b, a, a, b is a path of length six because (b, a), (a, c), (c, b), (b, a), (a, a), and (a, b) are all edges. We see that d, c is a path of length one, because (d, c) is an edge. Also c, b, a is a path of length two, because (c, b) and (b, a) are edges. All of (e, b), (b, a), (a, b), (b, a), (a, b), and (b, e) are edges, so e, b, a, b, a, b, e is a path of length six.

The two paths b, a, c, b, a, a, b and e, b, a, b, a, b, e are circuits because they begin and end at the same vertex. The paths a, b, e, d; c, b, a; and d, c are not circuits. ◀

The term *path* also applies to relations. Carrying over the definition from directed graphs to relations, there is a **path** from a to b in R if there is a sequence of elements $a, x_1, x_2, \ldots, x_{n-1}, b$ with $(a, x_1) \in R$, $(x_1, x_2) \in R, \ldots,$ and $(x_{n-1}, b) \in R$. Theorem 1 can be obtained from the definition of a path in a relation.

FIGURE 1 A Directed Graph.

THEOREM 1 Let R be a relation on a set A. There is a path of length n, where n is a positive integer, from a to b if and only if $(a, b) \in R^n$.

Proof: We will use mathematical induction. By definition, there is a path from a to b of length one if and only if $(a, b) \in R$, so the theorem is true when $n = 1$.

Assume that the theorem is true for the positive integer n. This is the inductive hypothesis. There is a path of length $n + 1$ from a to b if and only if there is an element $c \in A$ such that there is a path of length one from a to c, so $(a, c) \in R$, and a path of length n from c to b, that is, $(c, b) \in R^n$. Consequently, by the induction hypothesis, there is a path of length $n + 1$ from a to b if and only if there is an element c with $(a, c) \in R$ and $(c, b) \in R^n$. But there is such an element if and only if $(a, b) \in R^{n+1}$. Therefore, there is a path of length $n + 1$ from a to b if and only if $(a, b) \in R^{n+1}$. This completes the proof. ◁

Transitive Closures

We now show that finding the transitive closure of a relation is equivalent to determining which pairs of vertices in the associated directed graph are connected by a path. With this in mind, we define a new relation.

DEFINITION 2 Let R be a relation on a set A. The *connectivity relation* R^* consists of the pairs (a, b) such that there is a path of length at least one from a to b in R.

Because R^n consists of the pairs (a, b) such that there is a path of length n from a to b, it follows that R^* is the union of all the sets R^n. In other words,

$$R^* = \bigcup_{n=1}^{\infty} R^n.$$

The connectivity relation is useful in many models.

EXAMPLE 4 Let R be the relation on the set of all people in the world that contains (a, b) if a has met b. What is R^n, where n is a positive integer greater than one? What is R^*?

Solution: The relation R^2 contains (a, b) if there is a person c such that $(a, c) \in R$ and $(c, b) \in R$, that is, if there is a person c such that a has met c and c has met b. Similarly, R^n consists of those pairs (a, b) such that there are people $x_1, x_2, \ldots, x_{n-1}$ such that a has met x_1, x_1 has met $x_2, \ldots,$ and x_{n-1} has met b.

The relation R^* contains (a, b) if there is a sequence of people, starting with a and ending with b, such that each person in the sequence has met the next person in the sequence. (There are many interesting conjectures about R^*. Do you think that this connectivity relation includes the pair with you as the first element and the president of Mongolia as the second element? We will use graphs to model this application in Chapter 9.) ◀

EXAMPLE 5 Let R be the relation on the set of all subway stops in New York City that contains (a, b) if it is possible to travel from stop a to stop b without changing trains. What is R^n when n is a positive integer? What is R^*?

Solution: The relation R^n contains (a, b) if it is possible to travel from stop a to stop b by making $n - 1$ changes of trains. The relation R^* consists of the ordered pairs (a, b) where it is possible to travel from stop a to stop b making as many changes of trains as necessary. (The reader should verify these statements.) ◀

EXAMPLE 6 Let R be the relation on the set of all states in the United States that contains (a, b) if state a and state b have a common border. What is R^n, where n is a positive integer? What is R^*?

Solution: The relation R^n consists of the pairs (a, b), where it is possible to go from state a to state b by crossing exactly n state borders. R^* consists of the ordered pairs (a, b), where it is possible to go from state a to state b crossing as many borders as necessary. (The reader should verify these statements.) The only ordered pairs not in R^* are those containing states that are not connected to the continental United States (i.e., those pairs containing Alaska or Hawaii). ◀

Theorem 2 shows that the transitive closure of a relation and the associated connectivity relation are the same.

THEOREM 2 The transitive closure of a relation R equals the connectivity relation R^*.

Proof: Note that R^* contains R by definition. To show that R^* is the transitive closure of R we must also show that R^* is transitive and that $R^* \subseteq S$ whenever S is a transitive relation that contains R.

First, we show that R^* is transitive. If $(a, b) \in R^*$ and $(b, c) \in R^*$, then there are paths from a to b and from b to c in R. We obtain a path from a to c by starting with the path from a to b and following it with the path from b to c. Hence, $(a, c) \in R^*$. It follows that R^* is transitive.

Now suppose that S is a transitive relation containing R. Because S is transitive, S^n also is transitive (the reader should verify this) and $S^n \subseteq S$ (by Theorem 1 of Section 8.1). Furthermore, because

$$S^* = \bigcup_{k=1}^{\infty} S^k$$

and $S^k \subseteq S$, it follows that $S^* \subseteq S$. Now note that if $R \subseteq S$, then $R^* \subseteq S^*$, because any path in R is also a path in S. Consequently, $R^* \subseteq S^* \subseteq S$. Thus, any transitive relation that contains R must also contain R^*. Therefore, R^* is the transitive closure of R. ◁

Now that we know that the transitive closure equals the connectivity relation, we turn our attention to the problem of computing this relation. We do not need to examine arbitrarily long paths to determine whether there is a path between two vertices in a finite directed graph. As Lemma 1 shows, it is sufficient to examine paths containing no more than n edges, where n is the number of elements in the set.

LEMMA 1 Let A be a set with n elements, and let R be a relation on A. If there is a path of length at least one in R from a to b, then there is such a path with length not exceeding n. Moreover, when $a \neq b$, if there is a path of length at least one in R from a to b, then there is such a path with length not exceeding $n - 1$.

FIGURE 2 Producing a Path with Length Not Exceeding n.

Proof: Suppose there is a path from a to b in R. Let m be the length of the shortest such path. Suppose that $x_0, x_1, x_2, \ldots, x_{m-1}, x_m$, where $x_0 = a$ and $x_m = b$, is such a path.

Suppose that $a = b$ and that $m > n$, so that $m \geq n + 1$. By the pigeonhole principle, because there are n vertices in A, among the m vertices $x_0, x_1, \ldots, x_{m-1}$, at least two are equal (see Figure 2).

Suppose that $x_i = x_j$ with $0 \leq i < j \leq m - 1$. Then the path contains a circuit from x_i to itself. This circuit can be deleted from the path from a to b, leaving a path, namely, $x_0, x_1, \ldots, x_i, x_{j+1}, \ldots, x_{m-1}, x_m$, from a to b of shorter length. Hence, the path of shortest length must have length less than or equal to n.

The case where $a \neq b$ is left as an exercise for the reader. ◁

From Lemma 1, we see that the transitive closure of R is the union of R, R^2, $R^3, \ldots,$ and R^n. This follows because there is a path in R^* between two vertices if and only if there is a path between these vertices in R^i, for some positive integer i with $i \leq n$. Because

$$R^* = R \cup R^2 \cup R^3 \cup \cdots \cup R^n$$

and the zero–one matrix representing a union of relations is the join of the zero–one matrices of these relations, the zero–one matrix for the transitive closure is the join of the zero–one matrices of the first n powers of the zero–one matrix of R.

THEOREM 3 Let $\mathbf{M}_R$ be the zero–one matrix of the relation R on a set with n elements. Then the zero–one matrix of the transitive closure R^* is

$$\mathbf{M}_{R^*} = \mathbf{M}_R \vee \mathbf{M}_R^{[2]} \vee \mathbf{M}_R^{[3]} \vee \cdots \vee \mathbf{M}_R^{[n]}.$$

EXAMPLE 7 Find the zero–one matrix of the transitive closure of the relation R where

$$\mathbf{M}_R = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

Solution: From Theorem 3, it follows that the zero–one matrix of R^* is

$$\mathbf{M}_{R^*} = \mathbf{M}_R \vee \mathbf{M}_R^{[2]} \vee \mathbf{M}_R^{[3]}.$$

Because

$$\mathbf{M}_R^{[2]} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{M}_R^{[3]} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix},$$

it follows that

$$\mathbf{M}_{R^*} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} \vee \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \vee \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}.$$

◀

Links

Theorem 3 can be used as a basis for an algorithm for computing the matrix of the relation R^*. To find this matrix, the successive Boolean powers of $\mathbf{M}_R$, up to the nth power, are computed. As each power is calculated, its join with the join of all smaller powers is formed. When this is done with the nth power, the matrix for R^* has been found. This procedure is displayed as Algorithm 1.

ALGORITHM 1 A Procedure for Computing the Transitive Closure.

procedure *transitive closure* ($\mathbf{M}_R$: zero–one $n \times n$ matrix)
$\mathbf{A} := \mathbf{M}_R$
$\mathbf{B} := \mathbf{A}$
for $i := 2$ **to** n
begin
  $\mathbf{A} := \mathbf{A} \odot \mathbf{M}_R$
  $\mathbf{B} := \mathbf{B} \vee \mathbf{A}$
end {$\mathbf{B}$ is the zero–one matrix for R^*}

We can easily find the number of bit operations used by Algorithm 1 to determine the transitive closure of a relation. Computing the Boolean powers $\mathbf{M}_R, \mathbf{M}_R^{[2]}, \ldots, \mathbf{M}_R^{[n]}$ requires that $n-1$ Boolean products of $n \times n$ zero–one matrices be found. Each of these Boolean products can be found using $n^2(2n-1)$ bit operations. Hence, these products can be computed using $n^2(2n-1)(n-1)$ bit operations.

To find $\mathbf{M}_{R^*}$ from the n Boolean powers of $\mathbf{M}_R$, $n-1$ joins of zero–one matrices need to be found. Computing each of these joins uses n^2 bit operations. Hence, $(n-1)n^2$ bit operations are used in this part of the computation. Therefore, when Algorithm 1 is used, the matrix of the transitive closure of a relation on a set with n elements can be found using $n^2(2n-1)(n-1) + (n-1)n^2 = 2n^3(n-1)$, which is $O(n^4)$ bit operations. The remainder of this section describes a more efficient algorithm for finding transitive closures.

Warshall's Algorithm

Links

Warshall's algorithm, named after Stephen Warshall, who described this algorithm in 1960, is an efficient method for computing the transitive closure of a relation. Algorithm 1 can find the transitive closure of a relation on a set with n elements using $2n^3(n-1)$ bit operations. However, the transitive closure can be found by Warshall's algorithm using only $2n^3$ bit operations.

Remark: Warshall's algorithm is sometimes called the Roy–Warshall algorithm, because Bernard Roy described this algorithm in 1959.

Suppose that R is a relation on a set with n elements. Let $v_1, v_2, \ldots, v_n$ be an arbitrary listing of these n elements. The concept of the **interior vertices** of a path is used in Warshall's algorithm. If $a, x_1, x_2, \ldots, x_{m-1}, b$ is a path, its interior vertices are $x_1, x_2, \ldots, x_{m-1}$, that is, all the vertices of the path that occur somewhere other than as the first and last vertices in the path. For instance, the interior vertices of a path a, c, d, f, g, h, b, j in a directed graph are c, d, f, g, h, and b. The interior vertices of a, c, d, a, f, b are c, d, a, and f. (Note that the first vertex in the path is not an interior vertex unless it is visited again by the path, except as the last vertex. Similarly, the last vertex in the path is not an interior vertex unless it was visited previously by the path, except as the first vertex.)

Warshall's algorithm is based on the construction of a sequence of zero–one matrices. These matrices are $\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_n$, where $\mathbf{W}_0 = \mathbf{M}_R$ is the zero–one matrix of this relation, and $\mathbf{W}_k = [w_{ij}^{(k)}]$, where $w_{ij}^{(k)} = 1$ if there is a path from v_i to v_j such that all the interior vertices of this path are in the set $\{v_1, v_2, \ldots, v_k\}$ (the first k vertices in the list) and is 0 otherwise. (The first and last vertices in the path may be outside the set of the first k vertices in the list.) Note that $\mathbf{W}_n = \mathbf{M}_{R^*}$, because the (i, j)th entry of $\mathbf{M}_{R^*}$ is 1 if and only if there is a path from v_i to v_j, with all interior vertices in the set $\{v_1, v_2, \ldots, v_n\}$ (but these are the only vertices in the directed graph). Example 8 illustrates what the matrix $\mathbf{W}_k$ represents.

EXAMPLE 8 Let R be the relation with directed graph shown in Figure 3. Let a, b, c, d be a listing of the elements of the set. Find the matrices $\mathbf{W}_0, \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3$, and $\mathbf{W}_4$. The matrix $\mathbf{W}_4$ is the transitive closure of R.

FIGURE 3 The Directed Graph of the Relation *R*.

Solution: Let $v_1 = a$, $v_2 = b$, $v_3 = c$, and $v_4 = d$. $\mathbf{W}_0$ is the matrix of the relation. Hence,

$$\mathbf{W}_0 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

$\mathbf{W}_1$ has 1 as its (i, j)th entry if there is a path from v_i to v_j that has only $v_1 = a$ as an interior vertex. Note that all paths of length one can still be used because they have no interior vertices.

STEPHEN WARSHALL (BORN 1935) Stephen Warshall, born in New York City, went to public school in Brooklyn. He attended Harvard University, receiving his degree in mathematics in 1956. He never received an advanced degree, because at that time no programs were available in his areas of interest. However, he took graduate courses at several different universities and contributed to the development of computer science and software engineering.

After graduating from Harvard, Warshall worked at ORO (Operation Research Office), which was set up by Johns Hopkins to do research and development for the U.S. Army. In 1958 he left ORO to take a position at a company called Technical Operations, where he helped build a research and development laboratory for military software projects. In 1961 he left Technical Operations to found Massachusetts Computer Associates. Later, this company became part of Applied Data Research (ADR). After the merger, Warshall sat on the board of directors of ADR and managed a variety of projects and organizations. He retired from ADR in 1982.

During his career Warshall carried out research and development in operating systems, compiler design, language design, and operations research. In the 1971–1972 academic year he presented lectures on software engineering at French universities. There is an interesting anecdote about his proof that the transitive closure algorithm, now known as Warshall's algorithm, is correct. He and a colleague at Technical Operations bet a bottle of rum on who first could determine whether this algorithm always works. Warshall came up with his proof overnight, winning the bet and the rum, which he shared with the loser of the bet. Because Warshall did not like sitting at a desk, he did much of his creative work in unconventional places, such as on a sailboat in the Indian Ocean or in a Greek lemon orchard.

Also, there is now an allowable path from b to d, namely, b, a, d. Hence,

$$\mathbf{W}_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

$\mathbf{W}_2$ has 1 as its (i, j)th entry if there is a path from v_i to v_j that has only $v_1 = a$ and/or $v_2 = b$ as its interior vertices, if any. Because there are no edges that have b as a terminal vertex, no new paths are obtained when we permit b to be an interior vertex. Hence, $\mathbf{W}_2 = \mathbf{W}_1$.

$\mathbf{W}_3$ has 1 as its (i, j)th entry if there is a path from v_i to v_j that has only $v_1 = a$, $v_2 = b$, and/or $v_3 = c$ as its interior vertices, if any. We now have paths from d to a, namely, d, c, a, and from d to d, namely, d, c, d. Hence,

$$\mathbf{W}_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}.$$

Finally, $\mathbf{W}_4$ has 1 as its (i, j)th entry if there is a path from v_i to v_j that has $v_1 = a$, $v_2 = b$, $v_3 = c$, and/or $v_4 = d$ as interior vertices, if any. Because these are all the vertices of the graph, this entry is 1 if and only if there is a path from v_i to v_j. Hence,

$$\mathbf{W}_4 = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}.$$

This last matrix, $\mathbf{W}_4$, is the matrix of the transitive closure. ◀

Warshall's algorithm computes $\mathbf{M}_{R^*}$ by efficiently computing $\mathbf{W}_0 = \mathbf{M}_R$, $\mathbf{W}_1$, $\mathbf{W}_2, \ldots,$ $\mathbf{W}_n = \mathbf{M}_{R^*}$. This observation shows that we can compute $\mathbf{W}_k$ directly from $\mathbf{W}_{k-1}$: There is a path from v_i to v_j with no vertices other than $v_1, v_2, \ldots, v_k$ as interior vertices if and only if either there is a path from v_i to v_j with its interior vertices among the first $k-1$ vertices in the list, or there are paths from v_i to v_k and from v_k to v_j that have interior vertices only among the first $k-1$ vertices in the list. That is, either a path from v_i to v_j already existed before v_k was permitted as an interior vertex, or allowing v_k as an interior vertex produces a path that goes from v_i to v_k and then from v_k to v_j. These two cases are shown in Figure 4.

The first type of path exists if and only if $w_{ij}^{[k-1]} = 1$, and the second type of path exists if and only if both $w_{ik}^{[k-1]}$ and $w_{kj}^{[k-1]}$ are 1. Hence, $w_{ij}^{[k]}$ is 1 if and only if either $w_{ij}^{[k-1]}$ is 1 or both $w_{ik}^{[k-1]}$ and $w_{kj}^{[k-1]}$ are 1. This gives us Lemma 2.

LEMMA 2 Let $\mathbf{W}_k = [w_{ij}^{[k]}]$ be the zero–one matrix that has a 1 in its (i, j)th position if and only if there is a path from v_i to v_j with interior vertices from the set $\{v_1, v_2, \ldots, v_k\}$. Then

$$w_{ij}^{[k]} = w_{ij}^{[k-1]} \vee (w_{ik}^{[k-1]} \wedge w_{kj}^{[k-1]}),$$

whenever i, j, and k are positive integers not exceeding n.

FIGURE 4 Adding v_k to the Set of Allowable Interior Vertices.

Lemma 2 gives us the means to compute efficiently the matrices $\mathbf{W}_k$, $k = 1, 2, \ldots, n$. We display the pseudocode for Warshall's algorithm, using Lemma 2, as Algorithm 2.

ALGORITHM 2 **Warshall Algorithm.**

```
procedure Warshall (M_R : n × n zero–one matrix)
W := M_R
for k := 1 to n
begin
    for i := 1 to n
    begin
        for j := 1 to n
        w_ij := w_ij ∨ (w_ik ∧ w_kj)
    end
end {W = [w_ij] is M_R*}
```

The computational complexity of Warshall's algorithm can easily be computed in terms of bit operations. To find the entry $w_{ij}^{[k]}$ from the entries $w_{ij}^{[k-1]}$, $w_{ik}^{[k-1]}$, and $w_{kj}^{[k-1]}$ using Lemma 2 requires two bit operations. To find all n^2 entries of $\mathbf{W}_k$ from those of $\mathbf{W}_{k-1}$ requires $2n^2$ bit operations. Because Warshall's algorithm begins with $\mathbf{W}_0 = \mathbf{M}_R$ and computes the sequence of n zero–one matrices $\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_n = \mathbf{M}_{R^*}$, the total number of bit operations used is $n \cdot 2n^2 = 2n^3$.

Exercises

1. Let R be the relation on the set $\{0, 1, 2, 3\}$ containing the ordered pairs $(0, 1)$, $(1, 1)$, $(1, 2)$, $(2, 0)$, $(2, 2)$, and $(3, 0)$. Find the
   a) reflexive closure of R. **b)** symmetric closure of R.
2. Let R be the relation $\{(a, b) \mid a \neq b\}$ on the set of integers. What is the reflexive closure of R?
3. Let R be the relation $\{(a, b) \mid a \text{ divides } b\}$ on the set of integers. What is the symmetric closure of R?
4. How can the directed graph representing the reflexive closure of a relation on a finite set be constructed from the directed graph of the relation?

In Exercises 5–7 draw the directed graph of the reflexive closure of the relations with the directed graph shown.

5.

6.

7.

8. How can the directed graph representing the symmetric closure of a relation on a finite set be constructed from the directed graph for this relation?

9. Find the directed graphs of the symmetric closures of the relations with directed graphs shown in Exercises 5–7.

10. Find the smallest relation containing the relation in Example 2 that is both reflexive and symmetric.

11. Find the directed graph of the smallest relation that is both reflexive and symmetric for each of the relations with directed graphs shown in Exercises 5–7.

12. Suppose that the relation R on the finite set A is represented by the matrix $\mathbf{M}_R$. Show that the matrix that represents the reflexive closure of R is $\mathbf{M}_R \vee \mathbf{I}_n$.

13. Suppose that the relation R on the finite set A is represented by the matrix $\mathbf{M}_R$. Show that the matrix that represents the symmetric closure of R is $\mathbf{M}_R \vee \mathbf{M}_R^t$.

14. Show that the closure of a relation R with respect to a property **P**, if it exists, is the intersection of all the relations with property **P** that contain R.

15. When is it possible to define the "irreflexive closure" of a relation R, that is, a relation that contains R, is irreflexive, and is contained in every irreflexive relation that contains R?

16. Determine whether these sequences of vertices are paths in this directed graph.

a) a, b, c, e
b) b, e, c, b, e
c) a, a, b, e, d, e
d) b, c, e, d, a, a, b
e) b, c, c, b, e, d, e, d
f) $a, a, b, b, c, c, b, e, d$

17. Find all circuits of length three in the directed graph in Exercise 16.

18. Determine whether there is a path in the directed graph in Exercise 16 beginning at the first vertex given and ending at the second vertex given.

a) a, b **b)** b, a **c)** b, b
d) a, e **e)** b, d **f)** c, d
g) d, d **h)** e, a **i)** e, c

19. Let R be the relation on the set $\{1, 2, 3, 4, 5\}$ containing the ordered pairs $(1, 3)$, $(2, 4)$, $(3, 1)$, $(3, 5)$, $(4, 3)$, $(5, 1)$, $(5, 2)$, and $(5, 4)$. Find

a) R^2. **b)** R^3. **c)** R^4.
d) R^5. **e)** R^6. **f)** R^*.

20. Let R be the relation that contains the pair (a, b) if a and b are cities such that there is a direct non-stop airline flight from a to b. When is (a, b) in

a) R^2? **b)** R^3? **c)** R^*?

21. Let R be the relation on the set of all students containing the ordered pair (a, b) if a and b are in at least one common class and $a \neq b$. When is (a, b) in

a) R^2? **b)** R^3? **c)** R^*?

22. Suppose that the relation R is reflexive. Show that R^* is reflexive.

23. Suppose that the relation R is symmetric. Show that R^* is symmetric.

24. Suppose that the relation R is irreflexive. Is the relation R^2 necessarily irreflexive?

25. Use Algorithm 1 to find the transitive closures of these relations on $\{1, 2, 3, 4\}$.

a) $\{(1, 2), (2, 1), (2, 3), (3, 4), (4, 1)\}$
b) $\{(2, 1), (2, 3), (3, 1), (3, 4), (4, 1), (4, 3)\}$
c) $\{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}$
d) $\{(1, 1), (1, 4), (2, 1), (2, 3), (3, 1), (3, 2), (3, 4), (4, 2)\}$

26. Use Algorithm 1 to find the transitive closures of these relations on $\{a, b, c, d, e\}$.

a) $\{(a, c), (b, d), (c, a), (d, b), (e, d)\}$
b) $\{(b, c), (b, e), (c, e), (d, a), (e, b), (e, c)\}$
c) $\{(a, b), (a, c), (a, e), (b, a), (b, c), (c, a), (c, b), (d, a), (e, d)\}$
d) $\{(a, e), (b, a), (b, d), (c, d), (d, a), (d, c), (e, a), (e, b), (e, c), (e, e)\}$

27. Use Warshall's algorithm to find the transitive closures of the relations in Exercise 25.

28. Use Warshall's algorithm to find the transitive closures of the relations in Exercise 26.

29. Find the smallest relation containing the relation $\{(1, 2), (1, 4), (3, 3), (4, 1)\}$ that is

a) reflexive and transitive.

b) symmetric and transitive.
c) reflexive, symmetric, and transitive.

30. Finish the proof of the case when $a \neq b$ in Lemma 1.

31. Algorithms have been devised that use $O(n^{2.8})$ bit operations to compute the Boolean product of two $n \times n$ zero–one matrices. Assuming that these algorithms can be used, give big-O estimates for the number of bit operations using Algorithm 1 and using Warshall's algorithm to find the transitive closure of a relation on a set with n elements.

32. Devise an algorithm using the concept of interior vertices in a path to find the length of the shortest path between two vertices in a directed graph, if such a path exists.

33. Adapt Algorithm 1 to find the reflexive closure of the transitive closure of a relation on a set with n elements.

34. Adapt Warshall's algorithm to find the reflexive closure of the transitive closure of a relation on a set with n elements.

35. Show that the closure with respect to the property **P** of the relation $R = \{(0, 0), (0, 1), (1, 1), (2, 2)\}$ on the set $\{0, 1, 2\}$ does not exist if **P** is the property
a) "is not reflexive."
b) "has an odd number of elements."

8.5 Equivalence Relations

Introduction

In some programming languages the names of variables can contain an unlimited number of characters. However, there is a limit on the number of characters that are checked when a compiler determines whether two variables are equal. For instance, in traditional C, only the first eight characters of a variable name are checked by the compiler. (These characters are uppercase or lowercase letters, digits, or underscores.) Consequently, the compiler considers strings longer than eight characters that agree in their first eight characters the same. Let R be the relation on the set of strings of characters such that sRt, where s and t are two strings, if s and t are at least eight characters long and the first eight characters of s and t agree, or $s = t$. It is easy to see that R is reflexive, symmetric, and transitive. Moreover, R divides the set of all strings into classes, where all strings in a particular class are considered the same by a compiler for traditional C.

The integers a and b are related by the "congruence modulo 4" relation when 4 divides $a - b$. We will show later that this relation is reflexive, symmetric, and transitive. It is not hard to see that a is related to b if and only if a and b have the same remainder when divided by 4. It follows that this relation splits the set of integers into four different classes. When we care only what remainder an integer leaves when it is divided by 4, we need only know which class it is in, not its particular value.

These two relations, R and congruence modulo 4, are examples of equivalence relations, namely, relations that are reflexive, symmetric, and transitive. In this section we will show that such relations split sets into disjoint classes of equivalent elements. Equivalence relations arise whenever we care only whether an element of a set is in a certain class of elements, instead of caring about its particular identity.

Equivalence Relations

Links

In this section we will study relations with a particular combination of properties that allows them to be used to relate objects that are similar in some way.

DEFINITION 1 A relation on a set A is called an *equivalence relation* if it is reflexive, symmetric, and transitive.

Equivalence relations are important throughout mathematics and computer science. One reason for this is that in an equivalence relation, when two elements are related it makes sense to say they are equivalent.

DEFINITION 2 Two elements a and b that are related by an equivalence relation are called *equivalent*. The notation $a \sim b$ is often used to denote that a and b are equivalent elements with respect to a particular equivalence relation.

For the notion of equivalent elements to make sense, every element should be equivalent to itself, as the reflexive property guarantees for an equivalence relation. It makes sense to say that a and b are related (not just that a is related to b) by an equivalence relation, because when a is related to b, by the symmetric property, b is related to a. Furthermore, because an equivalence relation is transitive, if a and b are equivalent and b and c are equivalent, it follows that a and c are equivalent.

Examples 1–5 illustrate the notion of an equivalence relation.

EXAMPLE 1 Let R be the relation on the set of integers such that aRb if and only if $a = b$ or $a = -b$. In Section 8.1 we showed that R is reflexive, symmetric, and transitive. It follows that R is an equivalence relation. ◀

EXAMPLE 2 Let R be the relation on the set of real numbers such that aRb if and only if $a - b$ is an integer. Is R an equivalence relation?

Extra Examples

Solution: Because $a - a = 0$ is an integer for all real numbers a, aRa for all real numbers a. Hence, R is reflexive. Now suppose that aRb. Then $a - b$ is an integer, so $b - a$ is also an integer. Hence, bRa. It follows that R is symmetric. If aRb and bRc, then $a - b$ and $b - c$ are integers. Therefore, $a - c = (a - b) + (b - c)$ is also an integer. Hence, aRc. Thus, R is transitive. Consequently, R is an equivalence relation. ◀

One of the most widely used equivalence relations is congruence modulo m, where m is a positive integer greater than 1.

EXAMPLE 3 **Congruence Modulo m** Let m be a positive integer with $m > 1$. Show that the relation

$$R = \{(a, b) \mid a \equiv b \pmod{m}\}$$

is an equivalence relation on the set of integers.

Solution: Recall from Section 3.4 that $a \equiv b \pmod{m}$ if and only if m divides $a - b$. Note that $a - a = 0$ is divisible by m, because $0 = 0 \cdot m$. Hence, $a \equiv a \pmod{m}$, so congruence modulo m is reflexive. Now suppose that $a \equiv b \pmod{m}$. Then $a - b$ is divisible by m, so $a - b = km$, where k is an integer. It follows that $b - a = (-k)m$, so $b \equiv a \pmod{m}$. Hence, congruence modulo m is symmetric. Next, suppose that $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$. Then m divides both $a - b$ and $b - c$. Therefore, there are integers k and l with $a - b = km$ and $b - c = lm$. Adding these two equations shows that $a - c = (a - b) + (b - c) = km + lm = (k + l)m$. Thus, $a \equiv c \pmod{m}$. Therefore, congruence modulo m is transitive. It follows that congruence modulo m is an equivalence relation. ◀

EXAMPLE 4 Suppose that R is the relation on the set of strings of English letters such that aRb if and only if $l(a) = l(b)$, where $l(x)$ is the length of the string x. Is R an equivalence relation?

Solution: Because $l(a) = l(a)$, it follows that aRa whenever a is a string, so that R is reflexive. Next, suppose that aRb, so that $l(a) = l(b)$. Then bRa, because $l(b) = l(a)$. Hence, R is symmetric. Finally, suppose that aRb and bRc. Then $l(a) = l(b)$ and $l(b) = l(c)$. Hence, $l(a) = l(c)$, so aRc. Consequently, R is transitive. Because R is reflexive, symmetric, and transitive, it is an equivalence relation. ◀

EXAMPLE 5 Let n be a positive integer and S a set of strings. Suppose that R_n is the relation on S such that sR_nt if and only if $s = t$, or both s and t have at least n characters and the first n characters of s and t are the same. That is, a string of fewer than n characters is related only to itself; a string s with at least n characters is related to a string t if and only if t has at least n characters and t begins with the n characters at the start of s. For example, let $n = 3$ and let S be the set of all bit strings. Then sR_3t either when $s = t$ or both s and t are bit strings of length 3 or more that begin with the same three bits. For instance, $01R_301$ and $00111R_300101$, but $01 \not R_3 010$ and $01011 \not R_3 01110$.

Show that for every set S of strings and every positive integer n, R_n is an equivalence relation on S.

Solution: The relation R_n is reflexive because $s = s$, so that sR_ns whenever s is a string in S. If sR_nt, then either $s = t$ or s and t are both at least n characters long that begin with the same n characters. This means that tR_ns. We conclude that R_n is symmetric.

Now suppose that sR_nt and tR_nu. Then either $s = t$ or s and t are at least n characters long and s and t begin with the same n characters, and either $t = u$ or t and u are at least n characters long and t and u begin with the same n characters. From this, we can deduce that either $s = u$ or both s and u are n characters long and s and u begin with the same n characters (because in this case we know that s, t, and u are all at least n characters long and both s and u begin with the same n characters that t does). Consequently, R_n is transitive. It follows that R_n is an equivalence relation. ◀

In Examples 6 and 7 we look at two relations that are not equivalence relations.

EXAMPLE 6 Show that the "divides" relation on the set of positive integers in not an equivalence relation.

Solution: By Examples 9 and 15 in Section 8.1, we know that the "divides" relation is reflexive and transitive. However, by Example 12 in Section 8.1, we know that this relation is not symmetric (for instance, $2 \mid 4$ but $4 \nmid 2$). We conclude that the "divides" relation on the set of positive integers is not an equivalence relation. ◀

EXAMPLE 7 Let R be the relation on the set of real numbers such that xRy if and only if x and y are real numbers that differ by less than 1, that is $|x - y| < 1$. Show that R is not an equivalence relation.

Solution: R is reflexive because $|x - x| = 0 < 1$ whenever $x \in \mathbf{R}$. R is symmetric, for if xRy, where x and y are real numbers, then $|x - y| < 1$, which tells us that $|y - x| = |x - y| < 1$, so that yRx. However, R is not an equivalence relation because it is not transitive. Take $x = 2.8$, $y = 1.9$, and $z = 1.1$, so that $|x - y| = |2.8 - 1.9| = 0.9 < 1$, $|y - z| = |1.9 - 1.1| = 0.8 < 1$, but $|x - z| = |2.8 - 1.1| = 1.7 > 1$. That is, $2.8\,R\,1.9$, $1.9\,R\,1.1$, but $2.8 \not R\, 1.1$. ◀

Equivalence Classes

Let A be the set of all students in your school who graduated from high school. Consider the relation R on A that consists of all pairs (x, y), where x and y graduated from the same high school. Given a student x, we can form the set of all students equivalent to x with respect to R. This set consists of all students who graduated from the same high school as x did. This subset of A is called an equivalence class of the relation.

DEFINITION 3 Let R be an equivalence relation on a set A. The set of all elements that are related to an element a of A is called the *equivalence class* of a. The equivalence class of a with respect to R is denoted by $[a]_R$. When only one relation is under consideration, we can delete the subscript R and write $[a]$ for this equivalence class.

In other words, if R is an equivalence relation on a set A, the equivalence class of the element a is

$$[a]_R = \{s \mid (a, s) \in R\}.$$

If $b \in [a]_R$, then b is called a **representative** of this equivalence class. Any element of a class can be used as a representative of this class. That is, there is nothing special about the particular element chosen as the representative of the class.

EXAMPLE 8 What is the equivalence class of an integer for the equivalence relation of Example 1?

Solution: Because an integer is equivalent to itself and its negative in this equivalence relation, it follows that $[a] = \{-a, a\}$. This set contains two distinct integers unless $a = 0$. For instance, $[7] = \{-7, 7\}$, $[-5] = \{-5, 5\}$, and $[0] = \{0\}$. ◀

EXAMPLE 9 What are the equivalence classes of 0 and 1 for congruence modulo 4?

Solution: The equivalence class of 0 contains all integers a such that $a \equiv 0 \pmod 4$. The integers in this class are those divisible by 4. Hence, the equivalence class of 0 for this relation is

$$[0] = \{\ldots, -8, -4, 0, 4, 8, \ldots\}.$$

The equivalence class of 1 contains all the integers a such that $a \equiv 1 \pmod 4$. The integers in this class are those that have a remainder of 1 when divided by 4. Hence, the equivalence class of 1 for this relation is

$$[1] = \{\ldots, -7, -3, 1, 5, 9, \ldots\}.$$ ◀

In Example 9 the equivalence classes of 0 and 1 with respect to congruence modulo 4 were found. Example 9 can easily be generalized, replacing 4 with any positive integer m. The equivalence classes of the relation congruence modulo m are called the **congruence classes modulo** m. The congruence class of an integer a modulo m is denoted by $[a]_m$, so $[a]_m = \{\ldots, a - 2m, a - m, a, a + m, a + 2m, \ldots\}$. For instance, from Example 9 it follows that $[0]_4 = \{\ldots, -8, -4, 0, 4, 8, \ldots\}$ and $[1]_4 = \{\ldots, -7, -3, 1, 5, 9, \ldots\}$.

EXAMPLE 10 What is the equivalence class of the string 0111 with respect to the equivalence relation R_3 from Example 5 on the set of all bit strings? (Recall that sR_3t if and only if s and t are bit strings with $s = t$ or s and t are strings of at least three bits that start with the same three bits.)

Solution: The bit strings equivalent to 0111 are the bit strings with at least three bits that begin with 011. These are the bit strings 011, 0110, 0111, 01100, 01101, 01110, 01111, and so on. Consequently,

$$[011]_{R_3} = \{011, 0110, 0111, 01100, 01101, 01110, 01111, \ldots\}.$$ ◀

EXAMPLE 11 **Identifiers in the C Programming Language** In the C programming language, an **identifier** is the name of a variable, a function, or another type of entity. Each identifier is a nonempty string of characters where each character is a lowercase or an uppercase English letter, a digit, or an underscore, and the first character is a lowercase or an uppercase English letter. Identifiers can be any length. This allows developers to use as many characters as they want to name an entity, such as a variable. However, for compilers for some versions of C, there is a limit on the number of characters checked when two names are compared to see whether they refer to the same thing. For example, Standard C compilers consider two identifiers the same when they agree in their first 31 characters. Consequently, developers must be careful not to use identifiers with the same initial 31 characters for different things. We see that two identifiers are considered the same when they are related by the relation R_{31} in Example 5. Using Example 5, we know that R_{31}, on the set of all identifiers in Standard C, is an equivalence relation.

What are the equivalence classes of the each of the identifiers Number_of_tropical_storms, Number_of_named_tropical_storms, and Number_of_named_tropical_storms_in_the_Atlantic_in_2005?

Solution: Note that when an identifier is less than 31 characters long, by the definition of R_{31}, its equivalence class contains only itself. Because the identifier Number_of_tropical_storms is 25 characters long, its equivalence class contains exactly one element, namely, itself.

The identifier Number_of_named_tropical_storms is exactly 31 characters long. An identifier is equivalent to it when it starts with these same 31 characters. Consequently, every identifier at least 31 characters long that starts with Number_of_named_tropical_storms is equivalent to this identifier. It follows that the equivalence class of Number_of_named_tropical_storms is the set of all identifiers that begin with the 31 characters Number_of_named_tropical_storms.

An identifier is equivalent to the Number_of_named_tropical_storms_in_the_Atlantic_in_2005 if and only if it begins with its first 31 characters. Because these characters are Number_of_named_tropical_storms, we see that an identifier is equivalent to Number_of_named_tropical_storms_in_the_Atlantic_in_2005 if and only if it is equivalent to Number_of_named_tropical_storms. It follows that these last two identifiers have the same equivalence class. ◀

Equivalence Classes and Partitions

Let A be the set of students at your school who are majoring in exactly one subject, and let R be the relation on A consisting of pairs (x, y), where x and y are students with the same major. Then R is an equivalence relation, as the reader should verify. We can see that R splits all students in A into a collection of disjoint subsets, where each subset contains students with a specified major. For instance, one subset contains all students majoring (just) in computer science, and a second subset contains all students majoring in history. Furthermore, these subsets are equivalence classes of R. This example illustrates how the equivalence classes of an equivalence relation

partition a set into disjoint, nonempty subsets. We will make these notions more precise in the following discussion.

Let R be a relation on the set A. Theorem 1 shows that the equivalence classes of two elements of A are either identical or disjoint.

THEOREM 1 Let R be an equivalence relation on a set A. These statements for elements a and b of A are equivalent:

$(i)\ aRb \qquad (ii)\ [a] = [b] \qquad (iii)\ [a] \cap [b] \neq \emptyset$

Proof: We first show that (i) implies (ii). Assume that aRb. We will prove that $[a] = [b]$ by showing $[a] \subseteq [b]$ and $[b] \subseteq [a]$. Suppose $c \in [a]$. Then aRc. Because aRb and R is symmetric, we know that bRa. Furthermore, because R is transitive and bRa and aRc, it follows that bRc. Hence, $c \in [b]$. This shows that $[a] \subseteq [b]$. The proof that $[b] \subseteq [a]$ is similar; it is left as an exercise for the reader.

Second, we will show that (ii) implies (iii). Assume that $[a] = [b]$. It follows that $[a] \cap [b] \neq \emptyset$ because $[a]$ is nonempty (because $a \in [a]$ because R is reflexive).

Next, we will show that (iii) implies (i). Suppose that $[a] \cap [b] \neq \emptyset$. Then there is an element c with $c \in [a]$ and $c \in [b]$. In other words, aRc and bRc. By the symmetric property, cRb. Then by transitivity, because aRc and cRb, we have aRb.

Because (i) implies (ii), (ii) implies (iii), and (iii) implies (i), the three statements, (i), (ii), and (iii), are equivalent. ◁

We are now in a position to show how an equivalence relation *partitions* a set. Let R be an equivalence relation on a set A. The union of the equivalence classes of R is all of A, because an element a of A is in its own equivalence class, namely, $[a]_R$. In other words,

$$\bigcup_{a \in A} [a]_R = A.$$

In addition, from Theorem 1, it follows that these equivalence classes are either equal or disjoint, so

$$[a]_R \cap [b]_R = \emptyset,$$

when $[a]_R \neq [b]_R$.

These two observations show that the equivalence classes form a partition of A, because they split A into disjoint subsets. More precisely, a **partition** of a set S is a collection of disjoint nonempty subsets of S that have S as their union. In other words, the collection of subsets A_i, $i \in I$ (where I is an index set) forms a partition of S if and only if

$$A_i \neq \emptyset \text{ for } i \in I,$$

$$A_i \cap A_j = \emptyset \text{ when } i \neq j,$$

and

$$\bigcup_{i \in I} A_i = S.$$

(Here the notation $\bigcup_{i \in I} A_i$ represents the union of the sets A_i for all $i \in I$.) Figure 1 illustrates the concept of a partition of a set.

FIGURE 1 A Partition of a Set.

EXAMPLE 12 Suppose that $S = \{1, 2, 3, 4, 5, 6\}$. The collection of sets $A_1 = \{1, 2, 3\}$, $A_2 = \{4, 5\}$, and $A_3 = \{6\}$ forms a partition of S, because these sets are disjoint and their union is S. ◀

We have seen that the equivalence classes of an equivalence relation on a set form a partition of the set. The subsets in this partition are the equivalence classes. Conversely, every partition of a set can be used to form an equivalence relation. Two elements are equivalent with respect to this relation if and only if they are in the same subset of the partition.

To see this, assume that $\{A_i \mid i \in I\}$ is a partition on S. Let R be the relation on S consisting of the pair (x, y), where x and y belong to the same subset A_i in the partition. To show that R is an equivalence relation we must show that R is reflexive, symmetric, and transitive.

We see that $(a, a) \in R$ for every $a \in S$, because a is in the same subset as itself. Hence, R is reflexive. If $(a, b) \in R$, then b and a are in the same subset of the partition, so that $(b, a) \in R$ as well. Hence, R is symmetric. If $(a, b) \in R$ and $(b, c) \in R$, then a and b are in the same subset in the partition, X, and b and c are in the same subset of the partition, Y. Because the subsets of the partition are disjoint and b belongs to X and Y, it follows that $X = Y$. Consequently, a and c belong to the same subset of the partition, so $(a, c) \in R$. Thus, R is transitive.

It follows that R is an equivalence relation. The equivalence classes of R consist of subsets of S containing related elements, and by the definition of R, these are the subsets of the partition. Theorem 2 summarizes the connections we have established between equivalence relations and partitions.

THEOREM 2 Let R be an equivalence relation on a set S. Then the equivalence classes of R form a partition of S. Conversely, given a partition $\{A_i \mid i \in I\}$ of the set S, there is an equivalence relation R that has the sets A_i, $i \in I$, as its equivalence classes.

Example 13 shows how to construct an equivalence relation from a partition.

EXAMPLE 13 List the ordered pairs in the equivalence relation R produced by the partition $A_1 = \{1, 2, 3\}$, $A_2 = \{4, 5\}$, and $A_3 = \{6\}$ of $S = \{1, 2, 3, 4, 5, 6\}$, given in Example 12.

Solution: The subsets in the partition are the equivalence classes of R. The pair $(a, b) \in R$ if and only if a and b are in the same subset of the partition. The pairs $(1, 1)$, $(1, 2)$, $(1, 3)$, $(2, 1)$, $(2, 2)$, $(2, 3)$, $(3, 1)$, $(3, 2)$, and $(3, 3)$ belong to R because $A_1 = \{1, 2, 3\}$ is an equivalence class; the pairs $(4, 4)$, $(4, 5)$, $(5, 4)$, and $(5, 5)$ belong to R because $A_2 = \{4, 5\}$ is an equivalence class; and finally the pair $(6, 6)$ belongs to R because $\{6\}$ is an equivalence class. No pair other than those listed belongs to R. ◀

The congruence classes modulo m provide a useful illustration of Theorem 2. There are m different congruence classes modulo m, corresponding to the m different remainders

possible when an integer is divided by m. These m congruence classes are denoted by $[0]_m, [1]_m, \ldots, [m-1]_m$. They form a partition of the set of integers.

EXAMPLE 14 What are the sets in the partition of the integers arising from congruence modulo 4?

Solution: There are four congruence classes, corresponding to $[0]_4$, $[1]_4$, $[2]_4$, and $[3]_4$. They are the sets

$$[0]_4 = \{\ldots, -8, -4, 0, 4, 8, \ldots\},$$
$$[1]_4 = \{\ldots, -7, -3, 1, 5, 9, \ldots\},$$
$$[2]_4 = \{\ldots, -6, -2, 2, 6, 10, \ldots\},$$
$$[3]_4 = \{\ldots, -5, -1, 3, 7, 11, \ldots\}.$$

These congruence classes are disjoint, and every integer is in exactly one of them. In other words, as Theorem 2 says, these congruence classes form a partition. ◀

We now provide an example of a partition of the set of all strings arising from an equivalence relation on this set.

EXAMPLE 15 Let R_3 be the relation from Example 5. What are the sets in the partition of the set of all bit strings arising from the relation R_3 on the set of all bit strings? (Recall that $s R_3 t$, where s and t are bit strings, if $s = t$ or s and t are bit strings with at least three bits that agree in their first three bits.)

Solution: Note that every bit string of length less than three is equivalent only to itself. Hence $[\lambda]_{R_3} = \{\lambda\}$, $[0]_{R_3} = \{0\}$, $[1]_{R_3} = \{1\}$, $[00]_{R_3} = \{00\}$, $[01]_{R_3} = \{01\}$, $[10]_{R_3} = \{10\}$, and $[11]_{R_3} = \{11\}$. Note that every bit string of length three or more is equivalent to one of the eight bit strings 000, 001, 010, 011, 100, 101, 110, and 111. We have

$$[000]_{R_3} = \{000, 0000, 0001, 00000, 00001, 00010, 00011, \ldots\},$$

$$[001]_{R_3} = \{001, 0010, 0011, 00100, 00101, 00110, 00111, \ldots\},$$

$$[010]_{R_3} = \{010, 0100, 0101, 01000, 01001, 01010, 01011, \ldots\},$$

$$[011]_{R_3} = \{011, 0110, 0111, 01100, 01101, 01110, 01111, \ldots\},$$

$$[100]_{R_3} = \{100, 1000, 1001, 10000, 10001, 10010, 10011, \ldots\},$$

$$[101]_{R_3} = \{101, 1010, 1011, 10100, 10101, 10110, 10111, \ldots\},$$

$$[110]_{R_3} = \{110, 1100, 1101, 11000, 11001, 11010, 11011, \ldots\}, \text{ and}$$

$$[111]_{R_3} = \{111, 1110, 1111, 11100, 11101, 11110, 11111, \ldots\}.$$

These 15 equivalence classes are disjoint and every bit string is in exactly one of them. As Theorem 2 tells us, these equivalence classes partition the set of all bit strings. ◀

Exercises

1. Which of these relations on $\{0, 1, 2, 3\}$ are equivalence relations? Determine the properties of an equivalence relation that the others lack.
   a) $\{(0, 0), (1, 1), (2, 2), (3, 3)\}$
   b) $\{(0, 0), (0, 2), (2, 0), (2, 2), (2, 3), (3, 2), (3, 3)\}$
   c) $\{(0, 0), (1, 1), (1, 2), (2, 1), (2, 2), (3, 3)\}$

d) $\{(0, 0), (1, 1), (1, 3), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)\}$
e) $\{(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2), (2, 0), (2, 2), (3, 3)\}$

2. Which of these relations on the set of all people are equivalence relations? Determine the properties of an equivalence relation that the others lack.
a) $\{(a, b) \mid a$ and b are the same age$\}$
b) $\{(a, b) \mid a$ and b have the same parents$\}$
c) $\{(a, b) \mid a$ and b share a common parent$\}$
d) $\{(a, b) \mid a$ and b have met$\}$
e) $\{(a, b) \mid a$ and b speak a common language$\}$

3. Which of these relations on the set of all functions from $\mathbf{Z}$ to $\mathbf{Z}$ are equivalence relations? Determine the properties of an equivalence relation that the others lack.
a) $\{(f, g) \mid f(1) = g(1)\}$
b) $\{(f, g) \mid f(0) = g(0)$ or $f(1) = g(1)\}$
c) $\{(f, g) \mid f(x) - g(x) = 1$ for all $x \in \mathbf{Z}\}$
d) $\{(f, g) \mid$ for some $C \in \mathbf{Z}$, for all $x \in \mathbf{Z}$, $f(x) - g(x) = C\}$
e) $\{(f, g) \mid f(0) = g(1)$ and $f(1) = g(0)\}$

4. Define three equivalence relations on the set of students in your discrete mathematics class different from the relations discussed in the text. Determine the equivalence classes for each of these equivalence relations.

5. Define three equivalence relations on the set of buildings on a college campus. Determine the equivalence classes for each of these equivalence relations.

6. Define three equivalence relations on the set of classes offered at your school. Determine the equivalence classes for each of these equivalence relations.

7. Show that the relation of logical equivalence on the set of all compound propositions is an equivalence relation. What are the equivalence classes of $\mathbf{F}$ and of $\mathbf{T}$?

8. Let R be the relation on the set of all sets of real numbers such that $S\,R\,T$ if and only if S and T have the same cardinality. Show that R is an equivalence relation. What are the equivalence classes of the sets $\{0, 1, 2\}$ and $\mathbf{Z}$?

9. Suppose that A is a nonempty set, and f is a function that has A as its domain. Let R be the relation on A consisting of all ordered pairs (x, y) such that $f(x) = f(y)$.
a) Show that R is an equivalence relation on A.
b) What are the equivalence classes of R?

10. Suppose that A is a nonempty set and R is an equivalence relation on A. Show that there is a function f with A as its domain such that $(x, y) \in R$ if and only if $f(x) = f(y)$.

11. Show that the relation R consisting of all pairs (x, y) such that x and y are bit strings of length three or more that agree in their first three bits is an equivalence relation on the set of all bit strings of length three or more.

12. Show that the relation R consisting of all pairs (x, y) such that x and y are bit strings of length three or more that agree except perhaps in their first three bits is an equivalence relation on the set of all bit strings of length three or more.

13. Show that the relation R consisting of all pairs (x, y) such that x and y are bit strings that agree in their first and third bits is an equivalence relation on the set of all bit strings of length three or more.

14. Let R be the relation consisting of all pairs (x, y) such that x and y are strings of uppercase and lowercase English letters with the property that for every positive integer n, the nth characters in x and y are the same letter, either uppercase or lowercase. Show that R is an equivalence relation.

15. Let R be the relation on the set of ordered pairs of positive integers such that $((a, b), (c, d)) \in R$ if and only if $a + d = b + c$. Show that R is an equivalence relation.

16. Let R be the relation on the set of ordered pairs of positive integers such that $((a, b), (c, d)) \in R$ if and only if $ad = bc$. Show that R is an equivalence relation.

17. (*Requires calculus*)
a) Show that the relation R on the set of all differentiable functions from $\mathbf{R}$ to $\mathbf{R}$ consisting of all pairs (f, g) such that $f'(x) = g'(x)$ for all real numbers x is an equivalence relation.
b) Which functions are in the same equivalence class as the function $f(x) = x^2$?

18. (*Requires calculus*)
a) Let n be a positive integer. Show that the relation R on the set of all polynomials with real-valued coefficients consisting of all pairs (f, g) such that $f^{(n)}(x) = g^{(n)}(x)$ is an equivalence relation. [Here $f^{(n)}(x)$ is the nth derivative of $f(x)$.]
b) Which functions are in the same equivalence class as the function $f(x) = x^4$, where $n = 3$?

19. Let R be the relation on the set of all URLs (or Web addresses) such that $x\,R\,y$ if and only if the Web page at x is the same as the Web page at y. Show that R is an equivalence relation.

20. Let R be the relation on the set of all people who have visited a particular Web page such that $x\,R\,y$ if and only if person x and person y have followed the same set of links starting at this Web page (going from Web page to Web page until they stop using the Web). Show that R is an equivalence relation.

In Exercises 21–23 determine whether the relation with the directed graphs shown is an equivalence relation.

21.

22.

23.

24. Determine whether the relations represented by these zero–one matrices are equivalence relations.

a) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ **b)** $\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$ **c)** $\begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

25. Show that the relation R on the set of all bit strings such that $s\,R\,t$ if and only if s and t contain the same number of 1s is an equivalence relation.

26. What are the equivalence classes of the equivalence relations in Exercise 1?

27. What are the equivalence classes of the equivalence relations in Exercise 2?

28. What are the equivalence classes of the equivalence relations in Exercise 3?

29. What is the equivalence class of the bit string 011 for the equivalence relation in Exercise 25?

30. What are the equivalence classes of these bit strings for the equivalence relation in Exercise 11?
a) 010 **b)** 1011 **c)** 11111 **d)** 01010101

31. What are the equivalence classes of the bit strings in Exercise 30 for the equivalence relation from Exercise 11?

32. What are the equivalence classes of the bit strings in Exercise 30 for the equivalence relation from Exercise 12?

33. What are the equivalence classes of the bit strings in Exercise 30 for the equivalence relation R_4 from Example 5 on the set of all bit strings? (Recall that bit strings s and t are equivalent under R_4 if and only if they are equal or they are both at least four bits long and agree in their first four bits.)

34. What are the equivalence classes of the bit strings in Exercise 30 for the equivalence relation R_5 from Example 5 on the set of all bit strings? (Recall that bit strings s and t are equivalent under R_5 if and only if they are equal or they are both at least five bits long and agree in their first five bits.)

35. What is the congruence class $[n]_5$ (that is, the equivalence class of n with respect to congruence modulo 5) when n is
a) 2? **b)** 3? **c)** 6? **d)** -3?

36. What is the congruence class $[4]_m$ when m is
a) 2? **b)** 3? **c)** 6? **d)** 8?

37. Give a description of each of the congruence classes modulo 6.

38. What is the equivalence class of the strings with respect to the equivalence relation in Exercise 14?
a) *No* **b)** *Yes* **c)** *Help*

39. **a)** What is the equivalence class of $(1, 2)$ with respect to the equivalence relation in Exercise 15?
b) Give an interpretation of the equivalence classes for the equivalence relation R in Exercise 15. [*Hint:* Look at the difference $a - b$ corresponding to (a, b).]

40. **a)** What is the equivalence class of $(1, 2)$ with respect to the equivalence relation in Exercise 16?
b) Give an interpretation of the equivalence classes for the equivalence relation R in Exercise 16. [*Hint:* Look at the ratio a/b corresponding to (a, b).]

41. Which of these collections of subsets are partitions of $\{1, 2, 3, 4, 5, 6\}$?
a) $\{1, 2\}, \{2, 3, 4\}, \{4, 5, 6\}$ **b)** $\{1\}, \{2, 3, 6\}, \{4\}, \{5\}$
c) $\{2, 4, 6\}, \{1, 3, 5\}$ **d)** $\{1, 4, 5\}, \{2, 6\}$

42. Which of these collections of subsets are partitions of $\{-3, -2, -1, 0, 1, 2, 3\}$?
a) $\{-3, -1, 1, 3\}, \{-2, 0, 2\}$
b) $\{-3, -2, -1, 0\}, \{0, 1, 2, 3\}$
c) $\{-3, 3\}, \{-2, 2\}, \{-1, 1\}, \{0\}$
d) $\{-3, -2, 2, 3\}, \{-1, 1\}$

43. Which of these collections of subsets are partitions on the set of bit strings of length 8?
a) the set of bit strings that begin with 1, the set of bit strings that begin with 00, and the set of bit strings that begin with 01
b) the set of bit strings that contain the string 00, the set of bit strings that contain the string 01, the set of bit strings that contain the string 10, and the set of bit strings that contain the string 11
c) the set of bit strings that end with 00, the set of bit strings that end with 01, the set of bit strings that end with 10, and the set of bit strings that end with 11
d) the set of bit strings that end with 111, the set of bit strings that end with 011, and the set of bit strings that end with 00
e) the set of bit strings that have $3k$ ones, where k is a nonnegative integer; the set of bit strings that contain $3k + 1$ ones, where k is a nonnegative integer; and the set of bit strings that contain $3k + 2$ ones, where k is a nonnegative integer

44. Which of these collections of subsets are partitions of the set of integers?
a) the set of even integers and the set of odd integers
b) the set of positive integers and the set of negative integers
c) the set of integers divisible by 3, the set of integers leaving a remainder of 1 when divided by 3, and the set of integers leaving a remainder of 2 when divided by 3
d) the set of integers less than -100, the set of integers with absolute value not exceeding 100, and the set of integers greater than 100
e) the set of integers not divisible by 3, the set of even integers, and the set of integers that leave a remainder of 3 when divided by 6

45. Which of these are partitions of the set $\mathbf{Z} \times \mathbf{Z}$ of ordered pairs of integers?

a) the set of pairs (x, y), where x or y is odd; the set of pairs (x, y), where x is even; and the set of pairs (x, y), where y is even

b) the set of pairs (x, y), where both x and y are odd; the set of pairs (x, y), where exactly one of x and y is odd; and the set of pairs (x, y), where both x and y are even

c) the set of pairs (x, y), where x is positive; the set of pairs (x, y), where y is positive; and the set of pairs (x, y), where both x and y are negative

d) the set of pairs (x, y), where $3 \mid x$ and $3 \mid y$; the set of pairs (x, y), where $3 \mid x$ and $3 \nmid y$; the set of pairs (x, y), where $3 \nmid x$ and $3 \mid y$; and the set of pairs (x, y), where $3 \nmid x$ and $3 \nmid y$

e) the set of pairs (x, y), where $x > 0$ and $y > 0$; the set of pairs (x, y), where $x > 0$ and $y \leq 0$; the set of pairs (x, y), where $x \leq 0$ and $y > 0$; and the set of pairs (x, y), where $x \leq 0$ and $y \leq 0$

f) the set of pairs (x, y), where $x \neq 0$ and $y \neq 0$; the set of pairs (x, y), where $x = 0$ and $y \neq 0$; and the set of pairs (x, y), where $x \neq 0$ and $y = 0$

46. Which of these are partitions of the set of real numbers?

a) the negative real numbers, $\{0\}$, the positive real numbers

b) the set of irrational numbers, the set of rational numbers

c) the set of intervals $[k, k+1]$, $k = \ldots, -2, -1, 0, 1, 2, \ldots$

d) the set of intervals $(k, k+1)$, $k = \ldots, -2, -1, 0, 1, 2, \ldots$

e) the set of intervals $(k, k+1]$, $k = \ldots, -2, -1, 0, 1, 2, \ldots$

f) the sets $\{x + n \mid n \in \mathbf{Z}\}$ for all $x \in [0, 1)$

47. List the ordered pairs in the equivalence relations produced by these partitions of $\{0, 1, 2, 3, 4, 5\}$.

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

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

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

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

48. List the ordered pairs in the equivalence relations produced by these partitions of $\{a, b, c, d, e, f, g\}$.

a) $\{a, b\}, \{c, d\}, \{e, f, g\}$

b) $\{a\}, \{b\}, \{c, d\}, \{e, f\}, \{g\}$

c) $\{a, b, c, d\}, \{e, f, g\}$

d) $\{a, c, e, g\}, \{b, d\}, \{f\}$

A partition P_1 is called a **refinement** of the partition P_2 if every set in P_1 is a subset of one of the sets in P_2.

49. Show that the partition formed from congruence classes modulo 6 is a refinement of the partition formed from congruence classes modulo 3.

50. Show that the partition of the set of people living in the United States consisting of subsets of people living in the same county (or parish) and same state is a refinement of the partition consisting of subsets of people living in the same state.

51. Show that the partition of the set of bit strings of length 16 formed by equivalence classes of bit strings that agree on the last eight bits is a refinement of the partition formed from the equivalence classes of bit strings that agree on the last four bits.

In Exercises 52 and 53, R_n refers to the family of equivalence relations defined in Example 5. Recall that $s\ R_n\ t$, where s and t are two strings if $s = t$ or s and t are strings with at least n characters that agree in their first n characters.

52. Show that the partition of the set of all bit strings formed by equivalence classes of bit strings with respect to the equivalence relation R_4 is a refinement of the partition formed by equivalence classes of bit strings with respect to the equivalence relation R_3.

53. Show that the partition of the set of all identifiers in C formed by the equivalence classes of identifiers with respect to the equivalence relation R_{31} is a refinement of the partition formed by equivalence classes of identifiers with respect to the equivalence relation R_8. (Compilers for "old" C consider identifiers the same when their names agree in their first eight characters, while compilers in standard C consider identifiers the same when their names agree in their first 31 characters.)

54. Suppose that R_1 and R_2 are equivalence relations on a set A. Let P_1 and P_2 be the partitions that correspond to R_1 and R_2, respectively. Show that $R_1 \subseteq R_2$ if and only if P_1 is a refinement of P_2.

55. Find the smallest equivalence relation on the set $\{a, b, c, d, e\}$ containing the relation $\{(a, b), (a, c), (d, e)\}$.

56. Suppose that R_1 and R_2 are equivalence relations on the set S. Determine whether each of these combinations of R_1 and R_2 must be an equivalence relation.

a) $R_1 \cup R_2$ **b)** $R_1 \cap R_2$ **c)** $R_1 \oplus R_2$

57. Consider the equivalence relation from Example 3, namely, $R = \{(x, y) \mid x - y \text{ is an integer}\}$.

a) What is the equivalence class of 1 for this equivalence relation?

b) What is the equivalence class of 1/2 for this equivalence relation?

***58.** Each bead on a bracelet with three beads is either red, white, or blue, as illustrated in the figure shown.

Define the relation R between bracelets as: (B_1, B_2), where B_1 and B_2 are bracelets, belongs to R if and only

if B_2 can be obtained from B_1 by rotating it or rotating it and then reflecting it.

a) Show that R is an equivalence relation.

b) What are the equivalence classes of R?

***59.** Let R be the relation on the set of all colorings of the 2×2 checkerboard where each of the four squares is colored either red or blue so that (C_1, C_2), where C_1 and C_2 are 2×2 checkerboards with each of their four squares colored blue or red, belongs to R if and only if C_2 can be obtained from C_1 either by rotating the checkerboard or by rotating it and then reflecting it.

a) Show that R is an equivalence relation.

b) What are the equivalence classes of R?

60. a) Let R be the relation on the set of functions from $\mathbf{Z}^+$ to $\mathbf{Z}^+$ such that (f, g) belongs to R if and only if f is $\Theta(g)$ (see Section 3.2). Show that R is an equivalence relation.

b) Describe the equivalence class containing $f(n) = n^2$ for the equivalence relation of part (a).

61. Determine the number of different equivalence relations on a set with three elements by listing them.

62. Determine the number of different equivalence relations on a set with four elements by listing them.

***63.** Do we necessarily get an equivalence relation when we form the transitive closure of the symmetric closure of the reflexive closure of a relation?

***64.** Do we necessarily get an equivalence relation when we form the symmetric closure of the reflexive closure of the transitive closure of a relation?

65. Suppose we use Theorem 2 to form a partition P from an equivalence relation R. What is the equivalence relation R' that results if we use Theorem 2 again to form an equivalence relation from P?

66. Suppose we use Theorem 2 to form an equivalence relation R from a partition P. What is the partition P' that results if we use Theorem 2 again to form a partition from R?

67. Devise an algorithm to find the smallest equivalence relation containing a given relation.

***68.** Let $p(n)$ denote the number of different equivalence relations on a set with n elements (and by Theorem 2 the number of partitions of a set with n elements). Show that $p(n)$ satisfies the recurrence relation $p(n) = \sum_{j=0}^{n-1} C(n-1, j)p(n-j-1)$ and the initial condition $p(0) = 1$. (*Note:* The numbers $p(n)$ are called **Bell numbers** after the American mathematician E. T. Bell.)

69. Use Exercise 68 to find the number of different equivalence relations on a set with n elements, where n is a positive integer not exceeding 10.

8.6 Partial Orderings

Introduction

We often use relations to order some or all of the elements of sets. For instance, we order words using the relation containing pairs of words (x, y), where x comes before y in the dictionary. We schedule projects using the relation consisting of pairs (x, y), where x and y are tasks in a project such that x must be completed before y begins. We order the set of integers using the relation containing the pairs (x, y), where x is less than y. When we add all of the pairs of the form (x, x) to these relations, we obtain a relation that is reflexive, antisymmetric, and transitive. These are properties that characterize relations used to order the elements of sets.

Links

DEFINITION 1 A relation R on a set S is called a *partial ordering* or *partial order* if it is reflexive, antisymmetric, and transitive. A set S together with a partial ordering R is called a *partially ordered set,* or *poset,* and is denoted by (S, R). Members of S are called *elements* of the poset.

We give examples of posets in Examples 1–3.

EXAMPLE 1 Show that the "greater than or equal" relation ($\geq$) is a partial ordering on the set of integers.

Extra Examples

Solution: Because $a \geq a$ for every integer a, $\geq$ is reflexive. If $a \geq b$ and $b \geq a$, then $a = b$. Hence, $\geq$ is antisymmetric. Finally, $\geq$ is transitive because $a \geq b$ and $b \geq c$ imply that $a \geq c$. It follows that $\geq$ is a partial ordering on the set of integers and $(\mathbf{Z}, \geq)$ is a poset. ◀

EXAMPLE 2 The divisibility relation $|$ is a partial ordering on the set of positive integers, because it is reflexive, antisymmetric, and transitive, as was shown in Section 8.1. We see that $(\mathbf{Z}^+, |)$ is a poset. Recall that ($\mathbf{Z}^+$ denotes the set of positive integers.) ◀

EXAMPLE 3 Show that the inclusion relation $\subseteq$ is a partial ordering on the power set of a set S.

Solution: Because $A \subseteq A$ whenever A is a subset of S, $\subseteq$ is reflexive. It is antisymmetric because $A \subseteq B$ and $B \subseteq A$ imply that $A = B$. Finally, $\subseteq$ is transitive, because $A \subseteq B$ and $B \subseteq C$ imply that $A \subseteq C$. Hence, $\subseteq$ is a partial ordering on $P(S)$, and $(P(S), \subseteq)$ is a poset. ◀

Example 4 illustrates a relation that is not a partial ordering.

EXAMPLE 4 Let R be the relation on the set of people such that $x R y$ if x and y are people and x is older than y. Show that R is not a partial ordering.

Extra Examples

Solution: Note that R is antisymmetric because if a person x is older than a person y, then y is not older than x. That is, if $x R y$, then $y \not R x$. The relation R is transitive because if person x is older than person y and y is older than person z, then x is older than z. That is, if $x R y$ and $y R z$, then $x R z$. However, R is not reflexive, because no person is older than himself or herself. That is, $x \not R x$ for all people x. It follows that R is not a partial ordering. ◀

In different posets different symbols such as $\leq$, $\subseteq$, and $|$, are used for a partial ordering. However, we need a symbol that we can use when we discuss the ordering relation in an arbitrary poset. Customarily, the notation $a \preccurlyeq b$ is used to denote that $(a, b) \in R$ in an arbitrary poset (S, R). This notation is used because the "less than or equal to" relation on the set of real numbers is the most familiar example of a partial ordering and the symbol $\preccurlyeq$ is similar to the $\leq$ symbol. (Note that the symbol $\preccurlyeq$ is used to denote the relation in *any* poset, not just the "less than or equals" relation.) The notation $a \prec b$ denotes that $a \preccurlyeq b$, but $a \neq b$. Also, we say "a is less than b" or "b is greater than a" if $a \prec b$.

When a and b are elements of the poset $(S, \preccurlyeq)$, it is not necessary that either $a \preccurlyeq b$ or $b \preccurlyeq a$. For instance, in $(P(\mathbf{Z}), \subseteq)$, $\{1, 2\}$ is not related to $\{1, 3\}$, and vice versa, because neither set is contained within the other. Similarly, in $(\mathbf{Z}^+, |)$, 2 is not related to 3 and 3 is not related to 2, because $2 \nmid 3$ and $3 \nmid 2$. This leads to Definition 2.

DEFINITION 2 The elements a and b of a poset $(S, \preccurlyeq)$ are called *comparable* if either $a \preccurlyeq b$ or $b \preccurlyeq a$. When a and b are elements of S such that neither $a \preccurlyeq b$ nor $b \preccurlyeq a$, a and b are called *incomparable*.

EXAMPLE 5 In the poset $(\mathbf{Z}^+, |)$, are the integers 3 and 9 comparable? Are 5 and 7 comparable?

Solution: The integers 3 and 9 are comparable, because $3 \mid 9$. The integers 5 and 7 are incomparable, because $5 \nmid 7$ and $7 \nmid 5$. ◀

The adjective "partial" is used to describe partial orderings because pairs of elements may be incomparable. When every two elements in the set are comparable, the relation is called a **total ordering.**

DEFINITION 3 If $(S, \preccurlyeq)$ is a poset and every two elements of S are comparable, S is called a *totally ordered* or *linearly ordered set,* and $\preccurlyeq$ is called a *total order* or a *linear order.* A totally ordered set is also called a *chain.*

EXAMPLE 6 The poset $(\mathbf{Z}, \leq)$ is totally ordered, because $a \leq b$ or $b \leq a$ whenever a and b are integers. ◀

EXAMPLE 7 The poset $(\mathbf{Z}^+, |)$ is not totally ordered because it contains elements that are incomparable, such as 5 and 7. ◀

In Chapter 4 we noted that $(\mathbf{Z}^+, \leq)$ is well-ordered, where $\leq$ is the usual "less than or equal to" relation. We now define well-ordered sets.

DEFINITION 4 $(S, \preccurlyeq)$ is a *well-ordered set* if it is a poset such that $\preccurlyeq$ is a total ordering and every nonempty subset of S has a least element.

EXAMPLE 8 The set of ordered pairs of positive integers, $\mathbf{Z}^+ \times \mathbf{Z}^+$, with $(a_1, a_2) \preccurlyeq (b_1, b_2)$ if $a_1 < b_1$, or if $a_1 = b_1$ and $a_2 \leq b_2$ (the lexicographic ordering), is a well-ordered set. The verification of this is left as an exercise at the end of this section. The set $\mathbf{Z}$, with the usual $\leq$ ordering, is not well-ordered because the set of negative integers, which is a subset of $\mathbf{Z}$, has no least element. ◀

In Section 4.3 we showed how to use the principle of well-ordered induction (there called generalized induction) to prove results about a well-ordered set. We now state and prove that this proof technique is valid.

THEOREM 1 **THE PRINCIPLE OF WELL-ORDERED INDUCTION** Suppose that S is a well-ordered set. Then $P(x)$ is true for all $x \in S$, if

INDUCTIVE STEP: For every $y \in S$, if $P(x)$ is true for all $x \in S$ with $x \prec y$, then $P(y)$ is true.

Proof: Suppose it is not the case that $P(x)$ is true for all $x \in S$. Then there is an element $y \in S$ such that $P(y)$ is false. Consequently, the set $A = \{x \in S \mid P(x) \text{ is false}\}$ is nonempty. Because S is well ordered, A has a least element a. By the choice of a as a least element of A, we know that $P(x)$ is true for all $x \in S$ with $x \prec a$. This implies by the inductive step $P(a)$ is true. This contradiction shows that $P(x)$ must be true for all $x \in S$. ◁

Remark: We do not need a basis step in a proof using the principle of well-ordered induction because if x_0 is the least element of a well ordered set, the inductive step tells us that $P(x_0)$ is true. This follows because there are no elements $x \in S$ with $x \prec x_0$, so we know (using a vacuous proof) that $P(x)$ is true for all $x \in S$ with $x \prec x_0$.

The principle of well-ordered induction is a versatile technique for proving results about well-ordered sets. Even when it is possible to use mathematical induction for the set of positive integers to prove a theorem, it may be simpler to use the principle of well-ordered induction, as

we saw in Examples 5 and 6 in Section 4.2, where we proved a result about the well-ordered set $(\mathbf{N} \times \mathbf{N}, \preccurlyeq)$ where $\preccurlyeq$ is lexicographic ordering on $\mathbf{N} \times \mathbf{N}$.

Lexicographic Order

The words in a dictionary are listed in alphabetic, or lexicographic, order, which is based on the ordering of the letters in the alphabet. This is a special case of an ordering of strings on a set constructed from a partial ordering on the set. We will show how this construction works in any poset.

First, we will show how to construct a partial ordering on the Cartesian product of two posets, $(A_1, \preccurlyeq_1)$ and $(A_2, \preccurlyeq_2)$. The **lexicographic ordering** $\preccurlyeq$ on $A_1 \times A_2$ is defined by specifying that one pair is less than a second pair if the first entry of the first pair is less than (in A_1) the first entry of the second pair, or if the first entries are equal, but the second entry of this pair is less than (in A_2) the second entry of the second pair. In other words, (a_1, a_2) is less than (b_1, b_2), that is,

$$(a_1, a_2) \prec (b_1, b_2),$$

either if $a_1 \prec_1 b_1$ or if both $a_1 = b_1$ and $a_2 \prec_2 b_2$.

We obtain a partial ordering $\preccurlyeq$ by adding equality to the ordering $\prec$ on $A_1 \times A_2$. The verification of this is left as an exercise.

EXAMPLE 9 Determine whether $(3, 5) \prec (4, 8)$, whether $(3, 8) \prec (4, 5)$, and whether $(4, 9) \prec (4, 11)$ in the poset $(\mathbf{Z} \times \mathbf{Z}, \preccurlyeq)$, where $\preccurlyeq$ is the lexicographic ordering constructed from the usual $\leq$ relation on $\mathbf{Z}$.

Solution: Because $3 < 4$, it follows that $(3, 5) \prec (4, 8)$ and that $(3, 8) \prec (4, 5)$. We have $(4, 9) \prec (4, 11)$, because the first entries of $(4, 9)$ and $(4, 11)$ are the same but $9 < 11$. ◀

In Figure 1 the ordered pairs in $\mathbf{Z}^+ \times \mathbf{Z}^+$ that are less than $(3, 4)$ are highlighted. A lexicographic ordering can be defined on the Cartesian product of n posets $(A_1, \preccurlyeq_1)$, $(A_2, \preccurlyeq_2), \ldots, (A_n, \preccurlyeq_n)$. Define the partial ordering $\preccurlyeq$ on $A_1 \times A_2 \times \cdots \times A_n$ by

$$(a_1, a_2, \ldots, a_n) \prec (b_1, b_2, \ldots, b_n)$$

if $a_1 \prec_1 b_1$, or if there is an integer $i > 0$ such that $a_1 = b_1, \ldots, a_i = b_i$, and $a_{i+1} \prec_{i+1} b_{i+1}$. In other words, one n-tuple is less than a second n-tuple if the entry of the first n-tuple in the first position where the two n-tuples disagree is less than the entry in that position in the second n-tuple.

EXAMPLE 10 Note that $(1, 2, 3, 5) \prec (1, 2, 4, 3)$, because the entries in the first two positions of these 4-tuples agree, but in the third position the entry in the first 4-tuple, 3, is less than that in the second 4-tuple, 4. (Here the ordering on 4-tuples is the lexicographic ordering that comes from the usual "less than or equals" relation on the set of integers.) ◀

We can now define lexicographic ordering of strings. Consider the strings $a_1a_2 \ldots a_m$ and $b_1b_2 \ldots b_n$ on a partially ordered set S. Suppose these strings are not equal. Let t be the minimum of m and n. The definition of lexicographic ordering is that the string $a_1a_2 \ldots a_m$ is less than

FIGURE 1 The Ordered Pairs Less Than (3, 4) in Lexicographic Order.

$b_1 b_2 \ldots b_n$ if and only if

$$(a_1, a_2, \ldots, a_t) \prec (b_1, b_2, \ldots, b_t), \text{ or}$$
$$(a_1, a_2, \ldots, a_t) = (b_1, b_2, \ldots, b_t) \text{ and } m < n,$$

where $\prec$ in this inequality represents the lexicographic ordering of S^t. In other words, to determine the ordering of two different strings, the longer string is truncated to the length of the shorter string, namely, to $t = \min(m, n)$ terms. Then the t-tuples made up of the first t terms of each string are compared using the lexicographic ordering on S^t. One string is less than another string if the t-tuple corresponding to the first string is less than the t-tuple of the second string, or if these two t-tuples are the same, but the second string is longer. The verification that this is a partial ordering is left as an exercise for the reader.

EXAMPLE 11 Consider the set of strings of lowercase English letters. Using the ordering of letters in the alphabet, a lexicographic ordering on the set of strings can be constructed. A string is less than a second string if the letter in the first string in the first position where the strings differ comes before the letter in the second string in this position, or if the first string and the second string agree in all positions, but the second string has more letters. This ordering is the same as that used in dictionaries. For example,

discreet $\prec$ *discrete*,

because these strings differ first in the seventh position, and $e \prec t$. Also,

discreet $\prec$ *discreetness*,

because the first eight letters agree, but the second string is longer. Furthermore,

discrete $\prec$ *discretion*,

because

discrete $\prec$ *discreti*. ◀

FIGURE 2 Constructing the Hasse Diagram for $(\{1, 2, 3, 4\}, \leq)$.

Hasse Diagrams

Many edges in the directed graph for a finite poset do not have to be shown because they must be present. For instance, consider the directed graph for the partial ordering $\{(a, b) \mid a \leq b\}$ on the set $\{1, 2, 3, 4\}$, shown in Figure 2(a). Because this relation is a partial ordering, it is reflexive, and its directed graph has loops at all vertices. Consequently, we do not have to show these loops because they must be present; in Figure 2(b) loops are not shown. Because a partial ordering is transitive, we do not have to show those edges that must be present because of transitivity. For example, in Figure 2(c) the edges $(1, 3)$, $(1, 4)$, and $(2, 4)$ are not shown because they must be present. If we assume that all edges are pointed "upward" (as they are drawn in the figure), we do not have to show the directions of the edges; Figure 2(c) does not show directions.

In general, we can represent a partial ordering on a finite set using this procedure: Start with the directed graph for this relation. Because a partial ordering is reflexive, a loop is present at every vertex. Remove these loops. Remove all edges that must be in the partial ordering because of the presence of other edges and transitivity. For instance, if (a, b) and (b, c) are in the partial ordering, remove the edge (a, c), because it must be present also. Furthermore, if (c, d) is also in the partial ordering, remove the edge (a, d), because it must be present also. Finally, arrange each edge so that its initial vertex is below its terminal vertex (as it is drawn on paper). Remove all the arrows on the directed edges, because all edges point "upward" toward their terminal vertex. (The edges left correspond to pairs in the covering relation of the poset. See the preamble to Exercise 28.)

These steps are well defined, and only a finite number of steps need to be carried out for a finite poset. When all the steps have been taken, the resulting diagram contains sufficient information to find the partial ordering. This diagram is called a **Hasse diagram,** named after the twentieth-century German mathematician Helmut Hasse.

Links

HELMUT HASSE (1898–1979) Helmut Hasse was born in Kassel, Germany. He served in the German navy after high school. He began his university studies at Göttingen University in 1918, moving in 1920 to Marburg University to study under the number theorist Kurt Hensel. During this time, Hasse made fundamental contributions to algebraic number theory. He became Hensel's successor at Marburg, later becoming director of the famous mathematical institute at Göttingen in 1934, and took a position at Hamburg University in 1950. Hasse served for 50 years as an editor of *Crelle's Journal,* a famous German mathematics periodical, taking over the job of chief editor in 1936 when the Nazis forced Hensel to resign. During World War II Hasse worked on applied mathematics research for the German navy. He was noted for the clarity and personal style of his lectures and was devoted both to number theory and to his students. (Hasse has been controversial for connections with the Nazi party. Investigations have shown he was a strong German nationalist but not an ardent Nazi.)

FIGURE 3 Constructing the Hasse Diagram of $(\{1, 2, 3, 4, 6, 8, 12\}, |)$.

EXAMPLE 12 Draw the Hasse diagram representing the partial ordering $\{(a, b) \mid a \text{ divides } b\}$ on $\{1, 2, 3, 4, 6, 8, 12\}$.

Solution: Begin with the digraph for this partial order, as shown in Figure 3(a). Remove all loops, as shown in Figure 3(b). Then delete all the edges implied by the transitive property. These are $(1, 4)$, $(1, 6)$, $(1, 8)$, $(1, 12)$, $(2, 8)$, $(2, 12)$, and $(3, 12)$. Arrange all edges to point upward, and delete all arrows to obtain the Hasse diagram. The resulting Hasse diagram is shown in Figure 3(c). ◀

EXAMPLE 13 Draw the Hasse diagram for the partial ordering $\{(A, B) \mid A \subseteq B\}$ on the power set $P(S)$ where $S = \{a, b, c\}$.

Solution: The Hasse diagram for this partial ordering is obtained from the associated digraph by deleting all the loops and all the edges that occur from transitivity, namely, $(\emptyset, \{a, b\})$, $(\emptyset, \{a, c\})$, $(\emptyset, \{b, c\})$, $(\emptyset, \{a, b, c\})$, $(\{a\}, \{a, b, c\})$, $(\{b\}, \{a, b, c\})$, and $(\{c\}, \{a, b, c\})$. Finally all edges point upward, and arrows are deleted. The resulting Hasse diagram is illustrated in Figure 4. ◀

Maximal and Minimal Elements

Elements of posets that have certain extremal properties are important for many applications. An element of a poset is called maximal if it is not less than any element of the poset. That is, a is **maximal** in the poset $(S, \preccurlyeq)$ if there is no $b \in S$ such that $a \prec b$. Similarly, an element of a poset is called minimal if it is not greater than any element of the poset. That is, a is **minimal** if there is no element $b \in S$ such that $b \prec a$. Maximal and minimal elements are easy to spot using a Hasse diagram. They are the "top" and "bottom" elements in the diagram.

EXAMPLE 14 Which elements of the poset $(\{2, 4, 5, 10, 12, 20, 25\}, |)$ are maximal, and which are minimal?

Solution: The Hasse diagram in Figure 5 for this poset shows that the maximal elements are 12, 20, and 25, and the minimal elements are 2 and 5. As this example shows, a poset can have more than one maximal element and more than one minimal element. ◀

FIGURE 4 The Hasse Diagram of $(P(\{a, b, c\}), \subseteq)$.

FIGURE 5 The Hasse Diagram of a Poset.

Sometimes there is an element in a poset that is greater than every other element. Such an element is called the greatest element. That is, a is the **greatest element** of the poset $(S, \preccurlyeq)$ if $b \preccurlyeq a$ for all $b \in S$. The greatest element is unique when it exists [see Exercise 40(a) at the end of this section]. Likewise, an element is called the least element if it is less than all the other elements in the poset. That is, a is the **least element** of $(S, \preccurlyeq)$ if $a \preccurlyeq b$ for all $b \in S$. The least element is unique when it exists [see Exercise 40(b) at the end of the section].

EXAMPLE 15 Determine whether the posets represented by each of the Hasse diagrams in Figure 6 have a greatest element and a least element.

Solution: The least element of the poset with Hasse diagram (a) is a. This poset has no greatest element. The poset with Hasse diagram (b) has neither a least nor a greatest element. The poset with Hasse diagram (c) has no least element. Its greatest element is d. The poset with Hasse diagram (d) has least element a and greatest element d. ◀

EXAMPLE 16 Let S be a set. Determine whether there is a greatest element and a least element in the poset $(P(S), \subseteq)$.

Solution: The least element is the empty set, because $\emptyset \subseteq T$ for any subset T of S. The set S is the greatest element in this poset, because $T \subseteq S$ whenever T is a subset of S. ◀

EXAMPLE 17 Is there a greatest element and a least element in the poset $(\mathbf{Z}^+, |)$?

Solution: The integer 1 is the least element because $1|n$ whenever n is a positive integer. Because there is no integer that is divisible by all positive integers, there is no greatest element. ◀

FIGURE 6 Hasse Diagrams of Four Posets.

Sometimes it is possible to find an element that is greater than or equal to all the elements in a subset A of a poset $(S, \preccurlyeq)$. If u is an element of S such that $a \preccurlyeq u$ for all elements $a \in A$, then u is called an **upper bound** of A. Likewise, there may be an element less than or equal to all the elements in A. If l is an element of S such that $l \preccurlyeq a$ for all elements $a \in A$, then l is called a **lower bound** of A.

EXAMPLE 18 Find the lower and upper bounds of the subsets $\{a, b, c\}$, $\{j, h\}$, and $\{a, c, d, f\}$ in the poset with the Hasse diagram shown in Figure 7.

FIGURE 7 The Hasse Diagram of a Poset.

Solution: The upper bounds of $\{a, b, c\}$ are e, f, j, and h, and its only lower bound is a. There are no upper bounds of $\{j, h\}$, and its lower bounds are a, b, c, d, e, and f. The upper bounds of $\{a, c, d, f\}$ are f, h, and j, and its lower bound is a. ◀

The element x is called the **least upper bound** of the subset A if x is an upper bound that is less than every other upper bound of A. Because there is only one such element, if it exists, it makes sense to call this element *the* least upper bound [see Exercise 42(a) at the end of this section]. That is, x is the least upper bound of A if $a \preccurlyeq x$ whenever $a \in A$, and $x \preccurlyeq z$ whenever z is an upper bound of A. Similarly, the element y is called the **greatest lower bound** of A if y is a lower bound of A and $z \preccurlyeq y$ whenever z is a lower bound of A. The greatest lower bound of A is unique if it exists [see Exercise 42(b) at the end of this section]. The greatest lower bound and least upper bound of a subset A are denoted by glb(A) and lub(A), respectively.

EXAMPLE 19 Find the greatest lower bound and the least upper bound of $\{b, d, g\}$, if they exist, in the poset shown in Figure 7.

Solution: The upper bounds of $\{b, d, g\}$ are g and h. Because $g \prec h$, g is the least upper bound. The lower bounds of $\{b, d, g\}$ are a and b. Because $a \prec b$, b is the greatest lower bound. ◀

EXAMPLE 20 Find the greatest lower bound and the least upper bound of the sets $\{3, 9, 12\}$ and $\{1, 2, 4, 5, 10\}$, if they exist, in the poset $(\mathbf{Z}^+, |)$.

Extra Examples

Solution: An integer is a lower bound of $\{3, 9, 12\}$ if 3, 9, and 12 are divisible by this integer. The only such integers are 1 and 3. Because $1 \mid 3$, 3 is the greatest lower bound of $\{3, 9, 12\}$. The only lower bound for the set $\{1, 2, 4, 5, 10\}$ with respect to $|$ is the element 1. Hence, 1 is the greatest lower bound for $\{1, 2, 4, 5, 10\}$.

An integer is an upper bound for $\{3, 9, 12\}$ if and only if it is divisible by 3, 9, and 12. The integers with this property are those divisible by the least common multiple of 3, 9, and 12, which is 36. Hence, 36 is the least upper bound of $\{3, 9, 12\}$. A positive integer is an upper bound for the set $\{1, 2, 4, 5, 10\}$ if and only if it is divisible by 1, 2, 4, 5, and 10. The integers with this property are those integers divisible by the least common multiple of these integers, which is 20. Hence, 20 is the least upper bound of $\{1, 2, 4, 5, 10\}$. ◀

Lattices

A partially ordered set in which every pair of elements has both a least upper bound and a greatest lower bound is called a **lattice.** Lattices have many special properties. Furthermore, lattices are used in many different applications such as models of information flow and play an important role in Boolean algebra.

FIGURE 8 Hasse Diagrams of Three Posets.

EXAMPLE 21 Determine whether the posets represented by each of the Hasse diagrams in Figure 8 are lattices.

Solution: The posets represented by the Hasse diagrams in (a) and (c) are both lattices because in each poset every pair of elements has both a least upper bound and a greatest lower bound, as the reader should verify. On the other hand, the poset with the Hasse diagram shown in (b) is not a lattice, because the elements b and c have no least upper bound. To see this, note that each of the elements d, e, and f is an upper bound, but none of these three elements precedes the other two with respect to the ordering of this poset. ◀

EXAMPLE 22 Is the poset $(\mathbf{Z}^+, |)$ a lattice?

Solution: Let a and b be two positive integers. The least upper bound and greatest lower bound of these two integers are the least common multiple and the greatest common divisor of these integers, respectively, as the reader should verify. It follows that this poset is a lattice. ◀

EXAMPLE 23 Determine whether the posets $(\{1, 2, 3, 4, 5\}, |)$ and $(\{1, 2, 4, 8, 16\}, |)$ are lattices.

Solution: Because 2 and 3 have no upper bounds in $(\{1, 2, 3, 4, 5\}, |)$, they certainly do not have a least upper bound. Hence, the first poset is not a lattice.

Every two elements of the second poset have both a least upper bound and a greatest lower bound. The least upper bound of two elements in this poset is the larger of the elements and the greatest lower bound of two elements is the smaller of the elements, as the reader should verify. Hence, this second poset is a lattice. ◀

EXAMPLE 24 Determine whether $(P(S), \subseteq)$ is a lattice where S is a set.

Solution: Let A and B be two subsets of S. The least upper bound and the greatest lower bound of A and B are $A \cup B$ and $A \cap B$, respectively, as the reader can show. Hence, $(P(S), \subseteq)$ is a lattice. ◀

EXAMPLE 25 **The Lattice Model of Information Flow** In many settings the flow of information from one person or computer program to another is restricted via security clearances. We can use a lattice model to represent different information flow policies. For example, one common information flow policy is the *multilevel security policy* used in government and military systems. Each piece of information is assigned to a security class, and each security class is represented by a pair (A, C) where A is an *authority level* and C is a *category*. People and computer programs are then allowed access to information from a specific restricted set of security classes.

Links

The typical authority levels used in the U.S. government are unclassified (0), confidential (1), secret (2), and top secret (3). Categories used in security classes are the subsets of a set of all *compartments* relevant to a particular area of interest. Each compartment represents a particular

subject area. For example, if the set of compartments is {*spies, moles, double agents*}, then there are eight different categories, one for each of the eight subsets of the set of compartments, such as {*spies, moles*}.

We can order security classes by specifying that $(A_1, C_1) \preccurlyeq (A_2, C_2)$ if and only if $A_1 \leq A_2$ and $C_1 \subseteq C_2$. Information is permitted to flow from security class (A_1, C_1) into security class (A_2, C_2) if and only if $(A_1, C_1) \preccurlyeq (A_2, C_2)$. For example, information is permitted to flow from the security class (*secret*, {*spies, moles*}) into the security class (*top secret*, {*spies, moles, double agents*}), whereas information is not allowed to flow from the security class (*top secret*, {*spies, moles*}) into either of the security classes (*secret*, {*spies, moles, double agents*}) or (*top secret*, {*spies*}).

We leave it to the reader (see Exercise 48 at the end of this section) to show that the set of all security classes with the ordering defined in this example forms a lattice. ◀

Topological Sorting

Links

Suppose that a project is made up of 20 different tasks. Some tasks can be completed only after others have been finished. How can an order be found for these tasks? To model this problem we set up a partial order on the set of tasks so that $a \prec b$ if and only if a and b are tasks where b cannot be started until a has been completed. To produce a schedule for the project, we need to produce an order for all 20 tasks that is compatible with this partial order. We will show how this can be done.

We begin with a definition. A total ordering $\preccurlyeq$ is said to be **compatible** with the partial ordering R if $a \preccurlyeq b$ whenever aRb. Constructing a compatible total ordering from a partial ordering is called **topological sorting.*** We will need to use Lemma 1.

LEMMA 1 Every finite nonempty poset $(S, \preccurlyeq)$ has at least one minimal element.

Proof: Choose an element a_0 of S. If a_0 is not minimal, then there is an element a_1 with $a_1 \prec a_0$. If a_1 is not minimal, there is an element a_2 with $a_2 \prec a_1$. Continue this process, so that if a_n is not minimal, there is an element a_{n+1} with $a_{n+1} \prec a_n$. Because there are only a finite number of elements in the poset, this process must end with a minimal element a_n. ◁

The topological sorting algorithm we will describe works for any finite nonempty poset. To define a total ordering on the poset $(A, \preccurlyeq)$, first choose a minimal element a_1; such an element exists by Lemma 1. Next, note that $(A - \{a_1\}, \preccurlyeq)$ is also a poset, as the reader should verify. (Here by $\preccurlyeq$ we mean the restriction of the original relation $\preccurlyeq$ on A to $A - \{a_1\}$.) If it is nonempty, choose a minimal element a_2 of this poset. Then remove a_2 as well, and if there are additional elements left, choose a minimal element a_3 in $A - \{a_1, a_2\}$. Continue this process by choosing a_{k+1} to be a minimal element in $A - \{a_1, a_2, \ldots, a_k\}$, as long as elements remain.

Because A is a finite set, this process must terminate. The end product is a sequence of elements $a_1, a_2, \ldots, a_n$. The desired total ordering $\preccurlyeq_t$ is defined by

$$a_1 \prec_t a_2 \prec_t \cdots \prec_t a_n.$$

*"Topological sorting" is terminology used by computer scientists; mathematicians use the terminology "linearization of a partial ordering" for the same thing. In mathematics, topology is the branch of geometry dealing with properties of geometric figures that hold for all figures that can be transformed into one another by continuous bijections. In computer science, a topology is any arrangement of objects that can be connected with edges.

This total ordering is compatible with the original partial ordering. To see this, note that if $b \prec c$ in the original partial ordering, c is chosen as the minimal element at a phase of the algorithm where b has already been removed, for otherwise c would not be a minimal element. Pseudocode for this topological sorting algorithm is shown in Algorithm 1.

ALGORITHM 1 Topological Sorting

procedure *topological sort* $((S, \preccurlyeq)$: finite poset)
$k := 1$
while $S \neq \emptyset$
begin
  $a_k :=$ a minimal element of S {such an element exists by Lemma 1}
  $S := S - \{a_k\}$
  $k := k + 1$
end $\{a_1, a_2, \ldots, a_n$ is a compatible total ordering of $S\}$

EXAMPLE 26 Find a compatible total ordering for the poset $(\{1, 2, 4, 5, 12, 20\}, |)$.

Solution: The first step is to choose a minimal element. This must be 1, because it is the only minimal element. Next, select a minimal element of $(\{2, 4, 5, 12, 20\}, |)$. There are two minimal elements in this poset, namely, 2 and 5. We select 5. The remaining elements are $\{2, 4, 12, 20\}$. The only minimal element at this stage is 2. Next, 4 is chosen because it is the only minimal element of $(\{4, 12, 20\}, |)$. Because both 12 and 20 are minimal elements of $(\{12, 20\}, |)$, either can be chosen next. We select 20, which leaves 12 as the last element left. This produces the total ordering

$$1 \prec 5 \prec 2 \prec 4 \prec 20 \prec 12.$$

The steps used by this sorting algorithm are displayed in Figure 9. ◀

Topological sorting has an application to the scheduling of projects.

FIGURE 9 A Topological Sort of $(\{1, 2, 4, 5, 12, 20\}, |)$.

FIGURE 10 The Hasse Diagram for Seven Tasks.

FIGURE 11 A Topological Sort of the Tasks.

EXAMPLE 27 A development project at a computer company requires the completion of seven tasks. Some of these tasks can be started only after other tasks are finished. A partial ordering on tasks is set up by considering task $X \prec$ task Y if task Y cannot be started until task X has been completed. The Hasse diagram for the seven tasks, with respect to this partial ordering, is shown in Figure 10. Find an order in which these tasks can be carried out to complete the project.

Solution: An ordering of the seven tasks can be obtained by performing a topological sort. The steps of a sort are illustrated in Figure 11. The result of this sort, $A \prec C \prec B \prec E \prec F \prec D \prec G$, gives one possible order for the tasks. ◀

Exercises

1. Which of these relations on $\{0, 1, 2, 3\}$ are partial orderings? Determine the properties of a partial ordering that the others lack.
   a) $\{(0, 0), (1, 1), (2, 2), (3, 3)\}$
   b) $\{(0, 0), (1, 1), (2, 0), (2, 2), (2, 3), (3, 2), (3, 3)\}$
   c) $\{(0, 0), (1, 1), (1, 2), (2, 2), (3, 3)\}$
   d) $\{(0, 0), (1, 1), (1, 2), (1, 3), (2, 2), (2, 3), (3, 3)\}$
   e) $\{(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2), (2, 0), (2, 2), (3, 3)\}$
2. Which of these relations on $\{0, 1, 2, 3\}$ are partial orderings? Determine the properties of a partial ordering that the others lack.
   a) $\{(0, 0), (2, 2), (3, 3)\}$
   b) $\{(0, 0), (1, 1), (2, 0), (2, 2), (2, 3), (3, 3)\}$
   c) $\{(0, 0), (1, 1), (1, 2), (2, 2), (3, 1), (3, 3)\}$
   d) $\{(0, 0), (1, 1), (1, 2), (1, 3), (2, 0), (2, 2), (2, 3), (3, 0), (3, 3)\}$
   e) $\{(0, 0), (0, 1), (0, 2), (0, 3), (1, 0), (1, 1), (1, 2), (1, 3), (2, 0), (2, 2), (3, 3)\}$
3. Is (S, R) a poset if S is the set of all people in the world and $(a, b) \in R$, where a and b are people, if
   a) a is taller than b?
   b) a is not taller than b?
   c) $a = b$ or a is an ancestor of b?
   d) a and b have a common friend?
4. Is (S, R) a poset if S is the set of all people in the world and $(a, b) \in R$, where a and b are people, if
   a) a is no shorter than b?
   b) a weighs more than b?
   c) $a = b$ or a is a descendant of b?
   d) a and b do not have a common friend?
5. Which of these are posets?
   a) $(\mathbf{Z}, =)$ **b)** $(\mathbf{Z}, \neq)$ **c)** $(\mathbf{Z}, \geq)$ **d)** $(\mathbf{Z}, \nmid)$
6. Which of these are posets?
   a) $(\mathbf{R}, =)$ **b)** $(\mathbf{R}, <)$ **c)** $(\mathbf{R}, \leq)$ **d)** $(\mathbf{R}, \neq)$
7. Determine whether the relations represented by these zero–one matrices are partial orders.

   a) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ **b)** $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

   c) $\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}$

8. Determine whether the relations represented by these zero–one matrices are partial orders.

a) $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ **b)** $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$

c) $\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{bmatrix}$

In Exercises 9–11 determine whether the relation with the directed graph shown is a partial order.

9.

10.

11.

12. Let (S, R) be a poset. Show that (S, R^{-1}) is also a poset, where R^{-1} is the inverse of R. The poset (S, R^{-1}) is called the **dual** of (S, R).

13. Find the duals of these posets.

a) $(\{0, 1, 2\}, \leq)$ **b)** $(\mathbf{Z}, \geq)$
c) $(P(\mathbf{Z}), \supseteq)$ **d)** $(\mathbf{Z}^+, |)$

14. Which of these pairs of elements are comparable in the poset $(\mathbf{Z}^+, |)$?

a) 5, 15 **b)** 6, 9 **c)** 8, 16 **d)** 7, 7

15. Find two incomparable elements in these posets.

a) $(P(\{0, 1, 2\}), \subseteq)$ **b)** $(\{1, 2, 4, 6, 8\}, |)$

16. Let $S = \{1, 2, 3, 4\}$. With respect to the lexicographic order based on the usual "less than" relation,

a) find all pairs in $S \times S$ less than $(2, 3)$.
b) find all pairs in $S \times S$ greater than $(3, 1)$.
c) draw the Hasse diagram of the poset $(S \times S, \preccurlyeq)$.

17. Find the lexicographic ordering of these n-tuples:

a) $(1, 1, 2), (1, 2, 1)$ **b)** $(0, 1, 2, 3), (0, 1, 3, 2)$
c) $(1, 0, 1, 0, 1), (0, 1, 1, 1, 0)$

18. Find the lexicographic ordering of these strings of lowercase English letters:

a) *quack, quick, quicksilver, quicksand, quacking*
b) *open, opener, opera, operand, opened*
c) *zoo, zero, zoom, zoology, zoological*

19. Find the lexicographic ordering of the bit strings 0, 01, 11, 001, 010, 011, 0001, and 0101 based on the ordering $0 < 1$.

20. Draw the Hasse diagram for the "greater than or equal to" relation on $\{0, 1, 2, 3, 4, 5\}$.

21. Draw the Hasse diagram for the "less than or equal to" relation on $\{0, 2, 5, 10, 11, 15\}$.

22. Draw the Hasse diagram for divisibility on the set

a) $\{1, 2, 3, 4, 5, 6\}$. **b)** $\{3, 5, 7, 11, 13, 16, 17\}$.
c) $\{2, 3, 5, 10, 11, 15, 25\}$. **d)** $\{1, 3, 9, 27, 81, 243\}$.

23. Draw the Hasse diagram for divisibility on the set

a) $\{1, 2, 3, 4, 5, 6, 7, 8\}$. **b)** $\{1, 2, 3, 5, 7, 11, 13\}$.
c) $\{1, 2, 3, 6, 12, 24, 36, 48\}$.
d) $\{1, 2, 4, 8, 16, 32, 64\}$.

24. Draw the Hasse diagram for inclusion on the set $P(S)$, where $S = \{a, b, c, d\}$.

In Exercises 25–27 list all ordered pairs in the partial ordering with the accompanying Hasse diagram.

25.

26.

27.

Let $(S, \preccurlyeq)$ be a poset. We say that an element $y \in S$ **covers** an element $x \in S$ if $x \prec y$ and there is no element $z \in S$ such that $x \prec z \prec y$. The set of pairs (x, y) such that y covers x is called the **covering relation** of $(S, \preccurlyeq)$.

28. What is the covering relation of the partial ordering $\{(a, b) \mid a \text{ divides } b\}$ on $\{1, 2, 3, 4, 6, 12\}$?

29. What is the covering relation of the partial ordering $\{(A, B) \mid A \subseteq B\}$ on the power set of S, where $S = \{a, b, c\}$?

30. Show that the pair (x, y) belongs to the covering relation of the finite poset $(S, \preccurlyeq)$ if and only if x is lower than y and there is an edge joining x and y in the Hasse diagram of this poset.

31. Show that a finite poset can be reconstructed from its covering relation. [*Hint:* Show that the poset is the reflexive transitive closure of its covering relation.]

32. Answer these questions for the partial order represented by this Hasse diagram.

a) Find the maximal elements.
b) Find the minimal elements.
c) Is there a greatest element?
d) Is there a least element?
e) Find all upper bounds of $\{a, b, c\}$.
f) Find the least upper bound of $\{a, b, c\}$, if it exists.
g) Find all lower bounds of $\{f, g, h\}$.
h) Find the greatest lower bound of $\{f, g, h\}$, if it exists.

33. Answer these questions for the poset $(\{3, 5, 9, 15, 24, 45\}, |)$.
a) Find the maximal elements.
b) Find the minimal elements.
c) Is there a greatest element?
d) Is there a least element?
e) Find all upper bounds of $\{3, 5\}$.
f) Find the least upper bound of $\{3, 5\}$, if it exists.
g) Find all lower bounds of $\{15, 45\}$.
h) Find the greatest lower bound of $\{15, 45\}$, if it exists.

34. Answer these questions for the poset $(\{2, 4, 6, 9, 12, 18, 27, 36, 48, 60, 72\}, |)$.
a) Find the maximal elements.
b) Find the minimal elements.
c) Is there a greatest element?
d) Is there a least element?
e) Find all upper bounds of $\{2, 9\}$.
f) Find the least upper bound of $\{2, 9\}$, if it exists.
g) Find all lower bounds of $\{60, 72\}$.
h) Find the greatest lower bound of $\{60, 72\}$, if it exists.

35. Answer these questions for the poset $(\{\{1\}, \{2\}, \{4\}, \{1, 2\}, \{1, 4\}, \{2, 4\}, \{3, 4\}, \{1, 3, 4\}, \{2, 3, 4\}\}, \subseteq)$.
a) Find the maximal elements.
b) Find the minimal elements.
c) Is there a greatest element?
d) Is there a least element?
e) Find all upper bounds of $\{\{2\}, \{4\}\}$.
f) Find the least upper bound of $\{\{2\}, \{4\}\}$, if it exists.
g) Find all lower bounds of $\{\{1, 3, 4\}, \{2, 3, 4\}\}$.
h) Find the greatest lower bound of $\{\{1, 3, 4\}, \{2, 3, 4\}\}$, if it exists.

36. Give a poset that has
a) a minimal element but no maximal element.
b) a maximal element but no minimal element.
c) neither a maximal nor a minimal element.

37. Show that lexicographic order is a partial ordering on the Cartesian product of two posets.

38. Show that lexicographic order is a partial ordering on the set of strings from a poset.

39. Suppose that $(S, \preccurlyeq_1)$ and $(T, \preccurlyeq_2)$ are posets. Show that $(S \times T, \preccurlyeq)$ is a poset where $(s, t) \preccurlyeq (u, v)$ if and only if $s \preccurlyeq_1 u$ and $t \preccurlyeq_2 v$.

40. a) Show that there is exactly one greatest element of a poset, if such an element exists.
b) Show that there is exactly one least element of a poset, if such an element exists.

41. a) Show that there is exactly one maximal element in a poset with a greatest element.
b) Show that there is exactly one minimal element in a poset with a least element.

42. a) Show that the least upper bound of a set in a poset is unique if it exists.
b) Show that the greatest lower bound of a set in a poset is unique if it exists.

43. Determine whether the posets with these Hasse diagrams are lattices.

44. Determine whether these posets are lattices.
a) $(\{1, 3, 6, 9, 12\}, |)$ b) $(\{1, 5, 25, 125\}, |)$
c) $(\mathbf{Z}, \geq)$
d) $(P(S), \supseteq)$, where $P(S)$ is the power set of a set S

45. Show that every nonempty finite subset of a lattice has a least upper bound and a greatest lower bound.

46. Show that if the poset (S, R) is a lattice then the dual poset (S, R^{-1}) is also a lattice.

47. In a company, the lattice model of information flow is used to control sensitive information with security classes represented by ordered pairs (A, C). Here A is an authority level, which may be nonproprietary (0), proprietary (1), restricted (2), or registered (3). A category C is a subset of the set of all projects $\{$*Cheetah, Impala, Puma*$\}$. (Names of animals are often used as code names for projects in companies.)
a) Is information permitted to flow from (*Proprietary*, $\{$*Cheetah, Puma*$\}$) into (*Restricted*, $\{$*Puma*$\}$)?
b) Is information permitted to flow from (*Restricted*, $\{$*Cheetah*$\}$) into (*Registered*, $\{$*Cheetah, Impala*$\}$)?
c) Into which classes is information from (*Proprietary*, $\{$*Cheetah, Puma*$\}$) permitted to flow?
d) From which classes is information permitted to flow into the security class (*Restricted*, $\{$*Impala, Puma*$\}$)?

48. Show that the set S of security classes (A, C) is a lattice, where A is a positive integer representing an authority class and C is a subset of a finite set of compartments, with $(A_1, C_1) \preccurlyeq (A_2, C_2)$ if and only if $A_1 \leq A_2$ and $C_1 \subseteq C_2$. [*Hint*: First show that $(S, \preccurlyeq)$ is a poset and then show that the least upper bound and greatest lower bound of (A_1, C_1) and (A_2, C_2) are $(\max(A_1, A_2), C_1 \cup C_2)$ and $(\min(A_1, A_2), C_1 \cap C_2)$, respectively.]

***49.** Show that the set of all partitions of a set S with the relation $P_1 \preccurlyeq P_2$ if the partition P_1 is a refinement of the partition P_2 is a lattice. (See the preamble to Exercise 49 of Section 8.5.)

50. Show that every totally ordered set is a lattice.

51. Show that every finite lattice has a least element and a greatest element.

52. Give an example of an infinite lattice with

a) neither a least nor a greatest element.
b) a least but not a greatest element.
c) a greatest but not a least element.
d) both a least and a greatest element.

53. Verify that $(\mathbf{Z}^+ \times \mathbf{Z}^+, \preccurlyeq)$ is a well-ordered set, where $\preccurlyeq$ is lexicographic order, as claimed in Example 8.

54. Determine whether each of these posets is well-ordered.

a) $(S, \leq)$, where $S = \{10, 11, 12, \ldots\}$
b) $(\mathbf{Q} \cap [0, 1], \leq)$ (the set of rational numbers between 0 and 1 inclusive)
c) $(S, \leq)$, where S is the set of positive rational numbers with denominators not exceeding 3
d) $(\mathbf{Z}^-, \geq)$, where $\mathbf{Z}^-$ is the set of negative integers

A poset $(R, \preccurlyeq)$ is **well-founded** if there is no infinite decreasing sequence of elements in the poset, that is, elements $x_1, x_2, \ldots, x_n$ such that $\cdots \prec x_n \prec \cdots \prec x_2 \prec x_1$. A poset $(R, \preccurlyeq)$ is **dense** if for all $x \in S$ and $y \in S$ with $x \prec y$, there is an element $z \in R$ such that $x \prec z \prec y$.

55. Show that the poset $(\mathbf{Z}, \preccurlyeq)$, where $x \prec y$ if and only if $|x| < |y|$ is well-founded but is not a totally ordered set.

56. Show that a dense poset with at least two elements that are comparable is not well-founded.

57. Show that the poset of rational numbers with the usual "less than or equal to" relation, $(\mathbf{Q}, \leq)$, is a dense poset.

***58.** Show that the set of strings of lowercase English letters with lexicographic order is neither well-founded nor dense.

59. Show that a poset is well-ordered if and only if it is totally ordered and well-founded.

60. Show that a finite nonempty poset has a maximal element.

61. Find a compatible total order for the poset with the Hasse diagram shown in Exercise 32.

62. Find a compatible total order for the divisibility relation on the set $\{1, 2, 3, 6, 8, 12, 24, 36\}$.

63. Find an order different from that constructed in Example 27 for completing the tasks in the development project.

64. Schedule the tasks needed to build a house, by specifying their order, if the Hasse diagram representing these tasks is as shown in the figure.

65. Find an ordering of the tasks of a software project if the Hasse diagram for the tasks of the project is as shown.

Key Terms and Results

TERMS

binary relation from A to B: a subset of $A \times B$

relation on A: a binary relation from A to itself (i.e., a subset of $A \times A$)

$S \circ R$: composite of R and S

R^{-1}: inverse relation of R

R^n: nth power of R

reflexive: a relation R on A is reflexive if $(a, a) \in R$ for all $a \in A$

symmetric: a relation R on A is symmetric if $(b, a) \in R$ whenever $(a, b) \in R$

antisymmetric: a relation R on A is antisymmetric if $a = b$ whenever $(a, b) \in R$ and $(b, a) \in R$

transitive: a relation R on A is transitive if $(a, b) \in R$ and $(b, c) \in R$ implies that $(a, c) \in R$

n-ary relation on $A_1, A_2, \ldots, A_n$: a subset of $A_1 \times A_2 \times \cdots \times A_n$

relational data model: a model for representing databases using n-ary relations

primary key: a domain of an n-ary relation such that an n-tuple is uniquely determined by its value for this domain

composite key: the Cartesian product of domains of an n-ary relation such that an n-tuple is uniquely determined by its values in these domains

selection operator: a function that selects the n-tuples in an n-ary relation that satisfy a specified condition

projection: a function that produces relations of smaller degree from an n-ary relation by deleting fields

join: a function that combines n-ary relations that agree on certain fields

directed graph or digraph: a set of elements called vertices and ordered pairs of these elements, called edges

loop: an edge of the form (a, a)

closure of a relation R with respect to a property P: the relation S (if it exists) that contains R, has property **P**, and is contained within any relation that contains R and has property **P**

path in a digraph: a sequence of edges $(a, x_1), (x_1, x_2), \ldots, (x_{n-2}, x_{n-1}), (x_{n-1}, b)$ such that the terminal vertex of each edge is the initial vertex of the succeeding edge in the sequence

circuit (or cycle) in a digraph: a path that begins and ends at the same vertex

R^* (connectivity relation): the relation consisting of those ordered pairs (a, b) such that there is a path from a to b

equivalence relation: a reflexive, symmetric, and transitive relation

equivalent: if R is an equivalence relation, a is equivalent to b if aRb

$[a]_R$ (equivalence class of a with respect to R): the set of all elements of A that are equivalent to a

$[a]_m$ (congruence class modulo m): the set of integers congruent to a modulo m

partition of a set S: a collection of pairwise disjoint nonempty subsets that have S as their union

partial ordering: a relation that is reflexive, antisymmetric, and transitive

poset (S, R): a set S and a partial ordering R on this set

comparable: the elements a and b in the poset $(A, \preccurlyeq)$ are comparable if $a \preccurlyeq b$ or $b \preccurlyeq a$

incomparable: elements in a poset that are not comparable

total (or linear) ordering: a partial ordering for which every pair of elements are comparable

totally (or linearly) ordered set: a poset with a total (or linear) ordering

well-ordered set: a poset $(S, \preccurlyeq)$, where $\preccurlyeq$ is a total order and every nonempty subset of S has a least element

lexicographic order: a partial ordering of Cartesian products or strings (see pages 569–570)

Hasse diagram: a graphical representation of a poset where loops and all edges resulting from the transitive property are not shown, and the direction of the edges is indicated by the position of the vertices

maximal element: an element of a poset that is not less than any other element of the poset

minimal element: an element of a poset that is not greater than any other element of the poset

greatest element: an element of a poset greater than or equal to all other elements in this set

least element: an element of a poset less than or equal to all other elements in this set

upper bound of a set: an element in a poset greater than all other elements in the set

lower bound of a set: an element in a poset less than all other elements in the set

least upper bound of a set: an upper bound of the set that is less than all other upper bounds

greatest lower bound of a set: a lower bound of the set that is greater than all other lower bounds

lattice: a partially ordered set in which every two elements have a greatest lower bound and a least upper bound

compatible total ordering for a partial ordering: a total ordering that contains the given partial ordering

topological sort: the construction of a total ordering compatible with a given partial ordering

RESULTS

The reflexive closure of a relation R on the set A equals $R \cup \Delta$, where $\Delta = \{(a, a) \mid a \in A\}$.

The symmetric closure of a relation R on the set A equals $R \cup R^{-1}$, where $R^{-1} = \{(b, a) \mid (a, b) \in R\}$.

The transitive closure of a relation equals the connectivity relation formed from this relation.

Warshall's algorithm for finding the transitive closure of a relation (see pages 550–553).

Let R be an equivalence relation. Then the following three statements are equivalent: (1) $a\ R\ b$; (2) $[a]_R \cap [b]_R \neq \emptyset$; (3) $[a]_R = [b]_R$.

The equivalence classes of an equivalence relation on a set A form a partition of A. Conversely, an equivalence relation can be constructed from any partition so that the equivalence classes are the subsets in the partition.

The principle of well-ordered induction.

The topological sorting algorithm (see pages 576–578).

Review Questions

1. **a)** What is a relation on a set?
   b) How many relations are there on a set with n elements?
2. **a)** What is a reflexive relation?
   b) What is a symmetric relation?
   c) What is an antisymmetric relation?
   d) What is a transitive relation?
3. Give an example of a relation on the set $\{1, 2, 3, 4\}$ that is
   a) reflexive, symmetric, and not transitive.
   b) not reflexive, symmetric, and transitive.
   c) reflexive, antisymmetric, and not transitive.
   d) reflexive, symmetric, and transitive.
   e) reflexive, antisymmetric, and transitive.

4. a) How many reflexive relations are there on a set with n elements?
   b) How many symmetric relations are there on a set with n elements?
   c) How many antisymmetric relations are there on a set with n elements?
5. a) Explain how an n-ary relation can be used to represent information about students at a university.
   b) How can the 5-ary relation containing names of students, their addresses, telephone numbers, majors, and grade point averages be used to form a 3-ary relation containing the names of students, their majors, and their grade point averages?
   c) How can the 4-ary relation containing names of students, their addresses, telephone numbers, and majors and the 4-ary relation containing names of students, their student numbers, majors, and numbers of credit hours be combined into a single n-ary relation?
6. a) Explain how to use a zero–one matrix to represent a relation on a finite set.
   b) Explain how to use the zero–one matrix representing a relation to determine whether the relation is reflexive, symmetric, and/or antisymmetric.
7. a) Explain how to use a directed graph to represent a relation on a finite set.
   b) Explain how to use the directed graph representing a relation to determine whether a relation is reflexive, symmetric, and/or antisymmetric.
8. a) Define the reflexive closure and the symmetric closure of a relation.
   b) How can you construct the reflexive closure of a relation?
   c) How can you construct the symmetric closure of a relation?
   d) Find the reflexive closure and the symmetric closure of the relation $\{(1, 2), (2, 3), (2, 4), (3, 1)\}$ on the set $\{1, 2, 3, 4\}$.
9. a) Define the transitive closure of a relation.
   b) Can the transitive closure of a relation be obtained by including all pairs (a, c) such that (a, b) and (b, c) belong to the relation?
   c) Describe two algorithms for finding the transitive closure of a relation.
   d) Find the transitive closure of the relation $\{(1,1), (1,3), (2,1), (2,3), (2,4), (3,2), (3,4), (4,1)\}$.
10. a) Define an equivalence relation.
    b) Which relations on the set $\{a, b, c, d\}$ are equivalence relations and contain (a, b) and (b, d)?
11. a) Show that congruence modulo m is an equivalence relation whenever m is a positive integer.
    b) Show that the relation $\{(a, b) \mid a \equiv \pm b \pmod 7\}$ is an equivalence relation on the set of integers.
12. a) What are the equivalence classes of an equivalence relation?
    b) What are the equivalence classes of the "congruent modulo 5" relation?
    c) What are the equivalence classes of the equivalence relation in Question 11(b)?
13. Explain the relationship between equivalence relations on a set and partitions of this set.
14. a) Define a partial ordering.
    b) Show that the divisibility relation on the set of positive integers is a partial order.
15. Explain how partial orderings on the sets A_1 and A_2 can be used to define a partial ordering on the set $A_1 \times A_2$.
16. a) Explain how to construct the Hasse diagram of a partial order on a finite set.
    b) Draw the Hasse diagram of the divisibility relation on the set $\{2, 3, 5, 9, 12, 15, 18\}$.
17. a) Define a maximal element of a poset and the greatest element of a poset.
    b) Give an example of a poset that has three maximal elements.
    c) Give an example of a poset with a greatest element.
18. a) Define a lattice.
    b) Give an example of a poset with five elements that is a lattice and an example of a poset with five elements that is not a lattice.
19. a) Show that every finite subset of a lattice has a greatest lower bound and a least upper bound.
    b) Show that every lattice with a finite number of elements has a least element and a greatest element.
20. a) Define a well-ordered set.
    b) Describe an algorithm for producing a well-ordered set from a partially ordered set.
    c) Explain how the algorithm from (b) can be used to order the tasks in a project if each task can be done only after one or more of the other tasks have been completed.

Supplementary Exercises

1. Let S be the set of all strings of English letters. Determine whether these relations are reflexive, irreflexive, symmetric, antisymmetric, and/or transitive.
   a) $R_1 = \{(a, b) \mid a$ and b have no letters in common$\}$
   b) $R_2 = \{(a, b) \mid a$ and b are not the same length$\}$
   c) $R_3 = \{(a, b) \mid a$ is longer than $b\}$
2. Construct a relation on the set $\{a, b, c, d\}$ that is
   a) reflexive, symmetric, but not transitive.
   b) irreflexive, symmetric, and transitive.
   c) irreflexive, antisymmetric, and not transitive.
   d) reflexive, neither symmetric nor antisymmetric, and transitive.
   e) neither reflexive, irreflexive, symmetric, antisymmetric, nor transitive.

3. Show that the relation R on $\mathbf{Z} \times \mathbf{Z}$ defined by $(a, b)\, R\, (c, d)$ if and only if $a + d = b + c$ is an equivalence relation.
4. Show that a subset of an antisymmetric relation is also antisymmetric.
5. Let R be a reflexive relation on a set A. Show that $R \subseteq R^2$.
6. Suppose that R_1 and R_2 are reflexive relations on a set A. Show that $R_1 \oplus R_2$ is irreflexive.
7. Suppose that R_1 and R_2 are reflexive relations on a set A. Is $R_1 \cap R_2$ also reflexive? Is $R_1 \cup R_2$ also reflexive?
8. Suppose that R is a symmetric relation on a set A. Is $\overline{R}$ also symmetric?
9. Let R_1 and R_2 be symmetric relations. Is $R_1 \cap R_2$ also symmetric? Is $R_1 \cup R_2$ also symmetric?
10. A relation R is called **circular** if aRb and bRc imply that cRa. Show that R is reflexive and circular if and only if it is an equivalence relation.
11. Show that a primary key in an n-ary relation is a primary key in any projection of this relation that contains this key as one of its fields.
12. Is the primary key in an n-ary relation also a primary key in a larger relation obtained by taking the join of this relation with a second relation?
13. Show that the reflexive closure of the symmetric closure of a relation is the same as the symmetric closure of its reflexive closure.
14. Let R be the relation on the set of all mathematicians that contains the ordered pair (a, b) if and only if a and b have written a paper together.
    a) Describe the relation R^2.
    b) Describe the relation R^*.
    c) The **Erdős number** of a mathematician is 1 if this mathematician wrote a paper with the prolific Hungarian mathematician Paul Erdős, it is 2 if this mathematician did not write a joint paper with Erdős but wrote a joint paper with someone who wrote a joint paper with Erdős, and so on (except that the Erdős number of Erdős himself is 0). Give a definition of the Erdős number in terms of paths in R.
15. a) Give an example to show that the transitive closure of the symmetric closure of a relation is not necessarily the same as the symmetric closure of the transitive closure of this relation.
    b) Show, however, that the transitive closure of the symmetric closure of a relation must contain the symmetric closure of the transitive closure of this relation.
16. a) Let S be the set of subroutines of a computer program. Define the relation R by $\mathbf{P}\,R\,\mathbf{Q}$ if subroutine $\mathbf{P}$ calls subroutine $\mathbf{Q}$ during its execution. Describe the transitive closure of R.
    b) For which subroutines $\mathbf{P}$ does $(\mathbf{P}, \mathbf{P})$ belong to the transitive closure of R?
    c) Describe the reflexive closure of the transitive closure of R.
17. Suppose that R and S are relations on a set A with $R \subseteq S$ such that the closures of R and S with respect to a property $\mathbf{P}$ both exist. Show that the closure of R with respect to $\mathbf{P}$ is a subset of the closure of S with respect to $\mathbf{P}$.
18. Show that the symmetric closure of the union of two relations is the union of their symmetric closures.

Links

PAUL ERDŐS (1913–1996) Paul Erdős, born in Budapest, Hungary, was the son of two high school mathematics teachers. He was a child prodigy; at age 3 he could multiply three-digit numbers in his head, and at 4 he discovered negative numbers on his own. Because his mother did not want to expose him to contagious diseases, he was mostly home-schooled. At 17 Erdős entered Eötvős University, graduating four years later with a Ph.D. in mathematics. After graduating he spent four years at Manchester, England, on a postdoctoral fellowship. In 1938 he went to the United States because of the difficult political situation in Hungary, especially for Jews. He spent much of his time in the United States, except for 1954 to 1962, when he was banned as part of the paranoia of the McCarthy era. He also spent considerable time in Israel.

Erdős made many significant contributions to combinatorics and to number theory. One of the discoveries of which he was most proud is his elementary proof (in the sense that it does not use any complex analysis) of the Prime Number Theorem, which provides an estimate for the number of primes not exceeding a fixed positive integer. He also participated in the modern development of the Ramsey theory.

Erdős traveled extensively throughout the world to work with other mathematicians, visiting conferences, universities, and research laboratories. He had no permanent home. He devoted himself almost entirely to mathematics, traveling from one mathematician to the next, proclaiming "My brain is open." Erdős was the author or coauthor of more than 1500 papers and had more than 500 coauthors. Copies of his articles are kept by Ron Graham, a famous discrete mathematician with whom he collaborated extensively and who took care of many of his worldly needs.

Erdős offered rewards, ranging from \$10 to \$10,000, for the solution of problems that he found particularly interesting, with the size of the reward depending on the difficulty of the problem. He paid out close to \$4000. Erdős had his own special language, using such terms as "epsilon" (child), "boss" (woman), "slave" (man), "captured" (married), "liberated" (divorced), "Supreme Fascist" (God), "Sam" (United States), and "Joe" (Soviet Union). Although he was curious about many things, he concentrated almost all his energy on mathematical research. He had no hobbies and no full-time job. He never married and apparently remained celibate. Erdős was extremely generous, donating much of the money he collected from prizes, awards, and stipends for scholarships and to worthwhile causes. He traveled extremely lightly and did not like having many material possessions.

*19. Devise an algorithm, based on the concept of interior vertices, that finds the length of the longest path between two vertices in a directed graph, or determines that there are arbitrarily long paths between these vertices.

20. Which of these are equivalence relations on the set of all people?

a) $\{(x, y) \mid x$ and y have the same sign of the zodiac$\}$
b) $\{(x, y) \mid x$ and y were born in the same year$\}$
c) $\{(x, y) \mid x$ and y have been in the same city$\}$

*21. How many different equivalence relations with exactly three different equivalence classes are there on a set with five elements?

22. Show that $\{(x, y) \mid x - y \in \mathbf{Q}\}$ is an equivalence relation on the set of real numbers, where $\mathbf{Q}$ denotes the set of rational numbers. What are $[1]$, $[\frac{1}{2}]$, and $[\pi]$?

23. Suppose that $P_1 = \{A_1, A_2, \ldots, A_m\}$ and $P_2 = \{B_1, B_2, \ldots, B_n\}$ are both partitions of the set S. Show that the collection of nonempty subsets of the form $A_i \cap B_j$ is a partition of S that is a refinement of both P_1 and P_2 (see the preamble to Exercise 49 of Section 8.5).

*24. Show that the transitive closure of the symmetric closure of the reflexive closure of a relation R is the smallest equivalence relation that contains R.

25. Let $\mathbf{R}(S)$ be the set of all relations on a set S. Define the relation $\preccurlyeq$ on $\mathbf{R}(S)$ by $R_1 \preccurlyeq R_2$ if $R_1 \subseteq R_2$, where R_1 and R_2 are relations on S. Show that $(\mathbf{R}(S), \preccurlyeq)$ is a poset.

26. Let $\mathbf{P}(S)$ be the set of all partitions of the set S. Define the relation $\preccurlyeq$ on $\mathbf{P}(S)$ by $P_1 \preccurlyeq P_2$ if P_1 is a refinement of P_2 (see Exercise 49 of Section 8.5). Show that $(\mathbf{P}(S), \preccurlyeq)$ is a poset.

27. Schedule the tasks needed to cook a Chinese meal by specifying their order, if the Hasse diagram representing these tasks is as shown here.

A subset of a poset such that every two elements of this subset are comparable is called a **chain.** A subset of a poset is called an **antichain** if every two elements of this subset are incomparable.

28. Find all chains in the posets with the Hasse diagrams shown in Exercises 25–27 in Section 8.6.

29. Find all antichains in the posets with the Hasse diagrams shown in Exercises 25–27 in Section 8.6.

30. Find an antichain with the greatest number of elements in the poset with the Hasse diagram of Exercise 32 in Section 8.6.

31. Show that every maximal chain in a finite poset $(S, \preccurlyeq)$ contains a minimal element of S. (A maximal chain is a chain that is not a subset of a larger chain.)

**32. Show that every finite poset can be partitioned into k chains, where k is the largest number of elements in an antichain in this poset.

*33. Show that in any group of $mn + 1$ people there is either a list of $m + 1$ people where a person in the list (except for the first person listed) is a descendant of the previous person on the list, or there are $n + 1$ people such that none of these people is a descendant of any of the other n people. [*Hint:* Use Exercise 32.]

Suppose that $(S, \preccurlyeq)$ is a well-founded partially ordered set. The *principle of well-founded induction* states that $P(x)$ is true for all $x \in S$ if $\forall x(\forall y(y \prec x \rightarrow P(y)) \rightarrow P(x))$.

34. Show that no separate basis case is needed for the principle of well-founded induction. That is, $P(u)$ is true for all minimal elements u in S if $\forall x(\forall y(y \prec x \rightarrow P(y)) \rightarrow P(x))$.

*35. Show that the principle of well-founded induction is valid.

A relation R on a set A is a **quasi-ordering** on A if R is reflexive and transitive.

36. Let R be the relation on the set of all functions from $\mathbf{Z}^+$ to $\mathbf{Z}^+$ such that (f, g) belongs to R if and only if f is $O(g)$. Show that R is a quasi-ordering.

37. Let R be a quasi-ordering on a set A. Show that $R \cap R^{-1}$ is an equivalence relation.

*38. Let R be a quasi-ordering and let S be the relation on the set of equivalence classes of $R \cap R^{-1}$ such that (C, D) belongs to S, where C and D are equivalence classes of R, if and only if there are elements c of C and d of D such that (c, d) belongs to R. Show that S is a partial ordering.

Let L be a lattice. Define the **meet** ($\wedge$) and **join** ($\vee$) operations by $x \wedge y = \text{glb}(x, y)$ and $x \vee y = \text{lub}(x, y)$.

39. Show that the following properties hold for all elements x, y, and z of a lattice L.

a) $x \wedge y = y \wedge x$ and $x \vee y = y \vee x$ **(commutative laws)**
b) $(x \wedge y) \wedge z = x \wedge (y \wedge z)$ and $(x \vee y) \vee z = x \vee (y \vee z)$ **(associative laws)**
c) $x \wedge (x \vee y) = x$ and $x \vee (x \wedge y) = x$ **(absorption laws)**
d) $x \wedge x = x$ and $x \vee x = x$ **(idempotent laws)**

40. Show that if x and y are elements of a lattice L, then $x \vee y = y$ if and only if $x \wedge y = x$.

A lattice L is **bounded** if it has both an **upper bound**, denoted by 1, such that $x \preccurlyeq 1$ for all $x \in L$ and a **lower bound**, denoted by 0, such that $0 \preccurlyeq x$ for all $x \in L$.

41. Show that if L is a bounded lattice with upper bound 1 and lower bound 0 then these properties hold for all elements $x \in L$.

a) $x \vee 1 = 1$ **b)** $x \wedge 1 = x$
c) $x \vee 0 = x$ **d)** $x \wedge 0 = 0$

42. Show that every finite lattice is bounded.

A lattice is called **distributive** if $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$ and $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ for all x, y, and z in L.

***43.** Give an example of a lattice that is not distributive.

44. Show that the lattice $(P(S), \subseteq)$ where $P(S)$ is the power set of a finite set S is distributive.

45. Is the lattice $(\mathbf{Z}^+, |)$ distributive?

The **complement** of an element a of a bounded lattice L with upper bound 1 and lower bound 0 is an element b such that $a \vee b = 1$ and $a \wedge b = 0$. Such a lattice is **complemented** if every element of the lattice has a complement.

46. Give an example of a finite lattice where at least one element has more than one complement and at least one element has no complement.

47. Show that the lattice $(P(S), \subseteq)$ where $P(S)$ is the power set of a finite set S is complemented.

***48.** Show that if L is a finite distributive lattice, then an element of L has at most one complement.

The game of Chomp, introduced in Example 12 in Section 1.7, can be generalized for play on any finite partially ordered set $(S, \preceq)$ with a least element a. In this game, a move consists of selecting an element x in S and removing x and all elements larger than it from S. The loser is the player who is forced to select the least element a.

49. Show that the game of Chomp with cookies arranged in an $m \times n$ rectangular grid, described in Example 12 in Section 1.7, is the same as the game of Chomp on the poset $(S, |)$, where S is the set of all positive integers that divide $p^{m-1}q^{n-1}$, where p and q are distinct primes.

50. Show that if $(S, \preceq)$ has a greatest element b, then a winning strategy for Chomp on this poset exists. [*Hint:* Generalize the argument in Example 12 in Section 1.7.]

Computer Projects

Write programs with these input and output.

1. Given the matrix representing a relation on a finite set, determine whether the relation is reflexive and/or irreflexive.

2. Given the matrix representing a relation on a finite set, determine whether the relation is symmetric and/or antisymmetric.

3. Given the matrix representing a relation on a finite set, determine whether the relation is transitive.

4. Given a positive integer n, display all the relations on a set with n elements.

***5.** Given a positive integer n, determine the number of transitive relations on a set with n elements.

***6.** Given a positive integer n, determine the number of equivalence relations on a set with n elements.

***7.** Given a positive integer n, display all the equivalence relations on the set of the n smallest positive integers.

8. Given an n-ary relation, find the projection of this relation when specified fields are deleted.

9. Given an m-ary relation and an n-ary relation, and a set of common fields, find the join of these relations with respect to these common fields.

10. Given the matrix representing a relation on a finite set, find the matrix representing the reflexive closure of this relation.

11. Given the matrix representing a relation on a finite set, find the matrix representing the symmetric closure of this relation.

12. Given the matrix representing a relation on a finite set, find the matrix representing the transitive closure of this relation by computing the join of the powers of the matrix representing the relation.

13. Given the matrix representing a relation on a finite set, find the matrix representing the transitive closure of this relation using Warshall's algorithm.

14. Given the matrix representing a relation on a finite set, find the matrix representing the smallest equivalence relation containing this relation.

15. Given a partial ordering on a finite set, find a total ordering compatible with it using topological sorting.

Computations and Explorations

Use a computational program or programs you have written to do these exercises.

1. Display all the different relations on a set with four elements.
2. Display all the different reflexive and symmetric relations on a set with six elements.
3. Display all the reflexive and transitive relations on a set with five elements.
4. Determine how many transitive relations there are on a set with n elements for all positive integers n with $n \leq 7$.
5. Find the transitive closure of a relation of your choice on a set with at least 20 elements. Either use a relation that corresponds to direct links in a particular transportation or communications network or use a randomly generated relation.
6. Compute the number of different equivalence relations on a set with n elements for all positive integers n not exceeding 20.
7. Display all the equivalence relations on a set with seven elements.

*8. Display all the partial orders on a set with five elements.

*9. Display all the lattices on a set with five elements.

Writing Projects

Respond to these questions with essays using outside sources.

1. Discuss the concept of a fuzzy relation. How are fuzzy relations used?
2. Describe the basic principles of relational databases, going beyond what was covered in Section 8.2. How widely used are relational databases as compared with other types of databases?
3. Look up the original papers by Warshall and by Roy (in French) in which they develop algorithms for finding transitive closures. Discuss their approaches. Why do you suppose that what we call Warshall's algorithm was discovered independently by more than one person?
4. Describe how equivalence classes can be used to define the rational numbers as classes of pairs of integers and how the basic arithmetic operations on rational numbers can be defined following this approach. (See Exercise 40 in Section 8.5.)
5. Explain how Helmut Hasse used what we now call Hasse diagrams.
6. Describe some of the mechanisms used to enforce information flow policies in computer operating systems.
7. Discuss the use of the Program Evaluation and Review Technique (PERT) to schedule the tasks of a large complicated project. How widely is PERT used?
8. Discuss the use of the Critical Path Method (CPM) to find the shortest time for the completion of a project. How widely is CPM used?
9. Discuss the concept of *duality* in a lattice. Explain how duality can be used to establish new results.
10. Explain what is meant by a *modular lattice.* Describe some of the properties of modular lattices and describe how modular lattices arise in the study of projective geometry.