

Preface

This textbook discusses selected applications of linear algebra. The presentation is suitable for students who have completed or are taking concurrently a standard sophomore-level course in linear algebra.

Topics are drawn from a wide variety of fields including business, economics, engineering, physics, geometry, approximation theory, ecology, sociology, demography, and genetics. Also included is a brief introduction to game theory, Markov chains, and graph theory. At the end of the text there is a three chapter minicourse in linear programming which can be covered in about six lectures.

With a few clearly-marked exceptions, each application is in its own *independent* chapter, so that chapters can be deleted or permuted freely to fit individual needs and interests. Each topic begins with a list of linear algebra prerequisites in order that a reader can tell in advance if he or she has sufficient background to read the chapter.

Since the topics vary considerably in difficulty, we have included a *subjective* rating of each topic — easy, moderate, more difficult. (See the *Guide for the Instructor* following this preface.) Our evaluation is based more on the intrinsic difficulty of the material, rather than the number of prerequisites; thus a topic requiring fewer mathematical prerequisites may be rated harder than one requiring more prerequisites.

Since our primary objective is to present applications of linear algebra, proofs are often omitted. We assume the reader has met the linear algebra prerequisites and whenever results from other fields are needed, they are stated precisely (with motivation where possible), but usually without proof.

Although this text was written to be used with Howard Anton's *Elementary Linear Algebra*, Second Edition, John Wiley and Sons, Inc., 1977, we have avoided specialized notation or terminology so that this book can be utilized in conjunction with any standard undergraduate text in linear algebra.

There are several possible ways to use this book:

- (a) as a supplement to a standard linear algebra text;
- (b) as a textbook for a follow-up course to linear algebra;
- (c) as part of a self-study enrichment program or an introduction to mathematical research.

In addition this text may serve as a source of topics for a mathematical modeling or computer programming course.

We would like to express our appreciation to Kathleen R. McCabe who typed the entire manuscript. Her patience and skill contributed greatly to the appearance of this text. Our thanks are also due to Charles Shuman who assisted with the exercises and examples. Finally, we thank the entire Wiley staff, especially Judy Hirsch and Gary Ostedt, for their encouragement and guidance.

Chris Rorres

Howard Anton

Contents

1	Constructing Curves and Surfaces through Specified Points	1
2	Graph Theory	9
3	Theory of Games	25
4	Markov Chains	39
5	Leontief Economic Models	53
6	Forest Management	67
7	Equilibrium Temperature Distributions	79
8	Some Applications in Genetics	93
9	Age-Specific Population Growth	109
10	Harvesting of Animal Populations	125
11	Least Squares Fitting to Data	137
12	A Least Squares Model for Human Hearing: Fourier Series	149
13	Linear Programming I: A Geometric Approach	163
14	Linear Programming II: Basic Concepts	181
15	Linear Programming III: The Simplex Method	201
	<i>Answers to Exercises</i>	221

Guide for the Instructor

The table below classifies the first twelve chapters according to difficulty as follows:

- Easy** — *the average student who has met the stated prerequisites should be able to read the material with no help from the instructor*
- Moderate** — *the average student who has met the stated prerequisites may require a little help from the instructor.*
- More Difficult** — *the average student who has met the stated prerequisites will probably need help from the instructor.*

	Chapter											
	1	2	3	4	5	6	7	8	9	10	11	12
Easy	•										•	
Moderate		•	•	•	•	•	•					•
More Difficult								•	•	•		

With the exception of Chapter 10, which depends on Chapter 9, the first twelve chapters are independent and can be presented in any order. Chapters 13, 14, and 15 provide an introduction to linear programming suitable for students at the sophomore level. This material does not depend on the rest of the text and can be presented as soon as the students have met the linear algebra prerequisites.

Constructing Curves and Surfaces through Specified Points ¹

A technique for using determinants to construct lines, circles, and general conic sections through specified points in the plane is described. The procedure is also used to pass planes and spheres in three-dimensional space through fixed points.

PREREQUISITES:

Linear systems
Determinants
Analytic geometry

INTRODUCTION

One of the fundamental results in the theory of Linear Algebra is the following:

A homogeneous linear system with as many equations as unknowns has a nontrivial solution if and only if the determinant of the system is zero.

In this chapter, we show how this result may be used to determine the equations of various curves and surfaces through specified points. We proceed immediately to some specific examples.

A LINE THROUGH TWO POINTS

Suppose we are given two distinct points in the plane, (x_1, y_1) and (x_2, y_2) . There is a unique line,

$$c_1x + c_2y + c_3 = 0, \quad (1.1)$$

which passes through these two points. Notice that c_1 , c_2 , and c_3 are not all zero, and that these coefficients are unique only up to a multiplicative constant.

Since (x_1, y_1) and (x_2, y_2) lie on the line, substituting them in (1.1) gives the two equations

$$c_1x_1 + c_2y_1 + c_3 = 0 \quad (1.2)$$

$$c_1x_2 + c_2y_2 + c_3 = 0. \quad (1.3)$$

The three equations, (1.1), (1.2), and (1.3), may be grouped together in system form as

$$x c_1 + y c_2 + c_3 = 0$$

$$x_1 c_1 + y_1 c_2 + c_3 = 0$$

$$x_2 c_1 + y_2 c_2 + c_3 = 0.$$

In this form, we have a homogeneous system of three equations for c_1 , c_2 , and c_3 . Since c_1 , c_2 , and c_3 are not all zero, this system has a nontrivial solution, and so the determinant of the system must be zero. That is,

$$\begin{vmatrix} x & y & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix} = 0. \quad (1.4)$$

Consequently, every point (x, y) on the line satisfies (1.4), and conversely, every point (x, y) which satisfies (1.4) lies on the line.

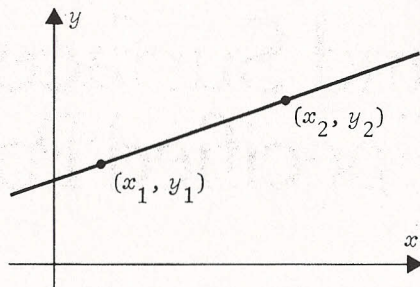


Figure 1.1

EXAMPLE 1.1 Find the equation of the line which passes through the two points $(2, 1)$ and $(3, 7)$.

SOLUTION Substitution of the coordinates of the two points into Eq. (1.4) gives

$$\begin{vmatrix} x & y & 1 \\ 2 & 1 & 1 \\ 3 & 7 & 1 \end{vmatrix} = 0.$$

The cofactor expansion of this determinant along the first row then gives:

$$-6x + y + 11 = 0.$$

A CIRCLE THROUGH THREE POINTS

Let us be given three distinct points in the plane, (x_1, y_1) , (x_2, y_2) , and (x_3, y_3) , not all lying on a straight line. From analytic geometry, we know that there is a unique circle, say

$$c_1(x^2 + y^2) + c_2x + c_3y + c_4 = 0, \quad (1.5)$$

which passes through them (Fig. 1.2). Substituting the coordinates of the three points into this equation gives

$$c_1(x_1^2 + y_1^2) + c_2x_1 + c_3y_1 + c_4 = 0 \quad (1.6)$$

$$c_1(x_2^2 + y_2^2) + c_2x_2 + c_3y_2 + c_4 = 0 \quad (1.7)$$

$$c_1(x_3^2 + y_3^2) + c_2x_3 + c_3y_3 + c_4 = 0. \quad (1.8)$$

As before, Eqs. (1.5) - (1.8) form a homogeneous linear system with a nontrivial solution for c_1 , c_2 , c_3 , and c_4 . Thus the determinant of this linear system is zero:

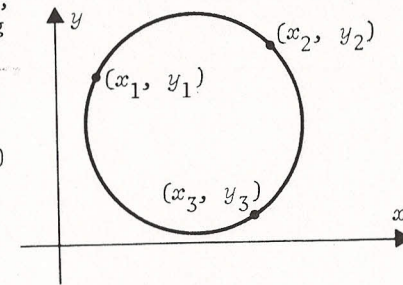


Figure 1.2

$$\begin{vmatrix} x^2 + y^2 & x & y & 1 \\ x_1^2 + y_1^2 & x_1 & y_1 & 1 \\ x_2^2 + y_2^2 & x_2 & y_2 & 1 \\ x_3^2 + y_3^2 & x_3 & y_3 & 1 \end{vmatrix} = 0. \tag{1.9}$$

This is a determinant form for the equation of the circle.

EXAMPLE 1.2 Find the equation of the circle which passes through the three points (1, 7), (6, 2), and (4, 6).

SOLUTION Substitution of the coordinates of the three points into Eq. (1.9) gives

$$\begin{vmatrix} x^2 + y^2 & x & y & 1 \\ 50 & 1 & 7 & 1 \\ 40 & 6 & 2 & 1 \\ 52 & 4 & 6 & 1 \end{vmatrix} = 0,$$

which reduces to

$$10(x^2 + y^2) - 20x - 40y - 200 = 0.$$

In standard form this is

$$(x - 1)^2 + (y - 2)^2 = 5^2.$$

Thus, the circle has center (1, 2) and radius 5.

A GENERAL CONIC SECTION THROUGH FIVE POINTS

The general equation of a conic section in the plane (a parabola, hyperbola, or ellipse, and degenerate forms of these three curves) is given by

$$c_1x^2 + c_2xy + c_3y^2 + c_4x + c_5y + c_6 = 0.$$

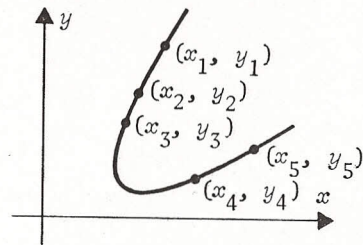


Figure 1.3

This form contains six coefficients, although only five are needed since we may divide through by any one of them which is not zero. Thus, only five coefficients must be determined, so that five distinct points in the plane are sufficient to determine the equation of the conic section (Fig. 1.3). As before, the equation may be put in determinant form (see Exercise 1.6):

$$\begin{vmatrix} x^2 & xy & y^2 & x & y & 1 \\ x_1^2 & x_1y_1 & y_1^2 & x_1 & y_1 & 1 \\ x_2^2 & x_2y_2 & y_2^2 & x_2 & y_2 & 1 \\ x_3^2 & x_3y_3 & y_3^2 & x_3 & y_3 & 1 \\ x_4^2 & x_4y_4 & y_4^2 & x_4 & y_4 & 1 \\ x_5^2 & x_5y_5 & y_5^2 & x_5 & y_5 & 1 \end{vmatrix} = 0. \tag{1.10}$$

EXAMPLE 1.3 An astronomer wants to determine the orbit of an asteroid about the sun. He sets up a Cartesian coordinate system in the plane of the orbit with the sun at the origin. Astronomical units of measurement are used along the axes. (1 astronomical unit = mean distance of earth to sun = 93 million miles.) By Kepler's first law, he knows that the orbit must be an ellipse. Consequently, he makes five observations of the asteroid at five different times, and finds five points along the orbit to be

$$(5.764, 0.648), (6.286, 1.202), (6.759, 1.823), (7.168, 2.526), (7.480, 3.360).$$

Find the equation of the orbit.

SOLUTION Substitution of the coordinates of the five given points into Eq. (1.10) gives:

$$\begin{vmatrix} x^2 & xy & y^2 & x & y & 1 \\ 33.224 & 3.735 & 0.420 & 5.764 & 0.648 & 1 \\ 39.514 & 7.556 & 1.445 & 6.286 & 1.202 & 1 \\ 45.684 & 12.322 & 3.323 & 6.759 & 1.823 & 1 \\ 51.380 & 18.106 & 6.381 & 7.168 & 2.526 & 1 \\ 55.950 & 25.133 & 11.290 & 7.480 & 3.360 & 1 \end{vmatrix} = 0.$$

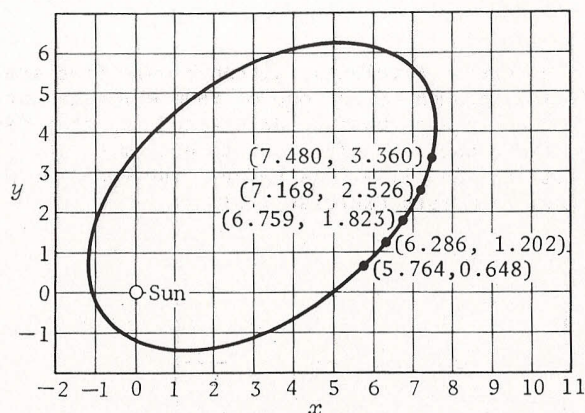


Figure 1.4

The cofactor expansion of this determinant along the first row reduces to

$$x^2 - 1.04xy + 1.30y^2 - 3.90x - 2.93y - 5.49 = 0.$$

Figure 1.4 is a diagram of the orbit, together with the five given points.

A PLANE THROUGH THREE POINTS

In Exercise 1.7 we ask the reader to show the following: The plane in 3-space with equation

$$c_1x + c_2y + c_3z + c_4 = 0$$

which passes through three noncollinear points (x_1, y_1, z_1) , (x_2, y_2, z_2) , and (x_3, y_3, z_3) is given by the determinant equation

$$\begin{vmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix} = 0. \quad (1.11)$$

EXAMPLE 1.4 The equation of the plane which passes through the three noncollinear points $(1, 1, 0)$, $(2, 0, -1)$, and $(2, 9, 2)$ is

$$\begin{vmatrix} x & y & z & 1 \\ 1 & 1 & 0 & 1 \\ 2 & 0 & -1 & 1 \\ 2 & 9 & 2 & 1 \end{vmatrix} = 0,$$

which reduces to

$$2x - y + 3z - 1 = 0.$$

A SPHERE THROUGH FOUR POINTS

In Exercise 1.8 we ask the reader to show the following: The sphere in 3-space with equation

$$c_1(x^2 + y^2 + z^2) + c_2x + c_3y + c_4z + c_5 = 0$$

which passes through four noncoplanar points (x_1, y_1, z_1) , (x_2, y_2, z_2) , (x_3, y_3, z_3) , and (x_4, y_4, z_4) is given by the following determinant equation:

$$\begin{vmatrix} x^2 + y^2 + z^2 & x & y & z & 1 \\ x_1^2 + y_1^2 + z_1^2 & x_1 & y_1 & z_1 & 1 \\ x_2^2 + y_2^2 + z_2^2 & x_2 & y_2 & z_2 & 1 \\ x_3^2 + y_3^2 + z_3^2 & x_3 & y_3 & z_3 & 1 \\ x_4^2 + y_4^2 + z_4^2 & x_4 & y_4 & z_4 & 1 \end{vmatrix} = 0. \quad (1.12)$$

EXAMPLE 1.5 The equation of the sphere which passes through the four points $(0, 3, 2)$, $(1, -1, 1)$, $(2, 1, 0)$, and $(-1, 1, 3)$ is

$$\begin{vmatrix} x^2 + y^2 + z^2 & x & y & z & 1 \\ 13 & 0 & 3 & 2 & 1 \\ 3 & 1 & -1 & 1 & 1 \\ 5 & 2 & 1 & 0 & 1 \\ 11 & -1 & 1 & 3 & 1 \end{vmatrix} = 0.$$

8 / Curves and Surfaces

This reduces to

$$x^2 + y^2 + z^2 - 4x - 2y - 6z + 5 = 0,$$

which in standard form is

$$(x - 2)^2 + (y - 1)^2 + (z - 3)^2 = 9.$$

EXERCISES

1.1 Find the equations of the lines which pass through the following points:

- (a) (1, -1), (2, 2)
- (b) (0, 1), (1, -1).

1.2 Find the equations of the circles which pass through the following points:

- (a) (2, 6), (2, 0), (5, 3)
- (b) (2, -2), (3, 5), (-4, 6).

1.3 Find the equation of the conic section which passes through the points (0, 0), (0, -1), (2, 0), (2, -5), and (4, -1).

1.4 Find the equations of the planes in 3-space which pass through the following points:

- (a) (1, 1, -3), (1, -1, 1), (0, -1, 2)
- (b) (2, 3, 1), (2, -1, -1), (1, 2, 1).

1.5 Find the equations of the spheres in 3-space which pass through the following points:

- (a) (1, 2, 3), (-1, 2, 1), (1, 0, 1), (1, 2, -1)
- (b) (0, 1, -2), (1, 3, 1), (2, -1, 0), (3, 1, -1).

1.6 Show that Eq. (1.10) is the equation of the conic section which passes through five given distinct points.

1.7 Show that Eq. (1.11) is the equation of the plane in 3-space which passes through three given noncollinear points.

1.8 Show that Eq. (1.12) is the equation of the sphere in 3-space which passes through four given noncoplanar points.

1.9 Find a determinant equation for the parabola of the form

$$a_1y + a_2x^2 + a_3x + a_4 = 0$$

which passes through three given noncollinear points in the plane.

Graph Theory

2

A matrix representation of relationships existing between members of a set is introduced. Through simple matrix arithmetic, the representation is used to analyze and clarify these relationships.

PREREQUISITES: Matrix multiplication and addition

INTRODUCTION

There are countless examples of sets consisting of a finite number of members in which some kind of relation exists between members of the set. For example, the set could consist of a collection of people, animals, countries, companies, sports teams or cities; and the relation between two members, A and B , from such a set could be that person A dominates person B , animal A feeds on animal B , country A militarily supports country B , company A sells its product to company B , sports team A consistently beats sports team B , or city A has a direct airline flight to city B .

In this chapter, we shall study the theory of directed graphs to mathematically model the types of sets and relations in the above examples.

DIRECTED GRAPHS

By a directed graph we mean a finite set of elements, P_1, P_2, \dots, P_n , together with a finite collection of ordered pairs, (P_i, P_j) , of distinct elements of this set, with no ordered pair being repeated. The elements of the set are called *vertices*, and the ordered pairs are called *directed edges*, of the directed graph. We use the notation $P_i \rightarrow P_j$ (read " P_i is connected to P_j ") to indicate that the directed edge (P_i, P_j) belongs to the directed graph. Graphically, we can visualize a directed graph (Fig. 2.1) by representing the vertices as points in the plane, and representing $P_i \rightarrow P_j$ by drawing a line or arc from vertex P_i to vertex P_j , with an arrow pointing from P_i to P_j . If both $P_i \rightarrow P_j$ and $P_j \rightarrow P_i$ hold (denoted $P_i \leftrightarrow P_j$), we draw a single line between P_i and P_j with two oppositely pointing arrows (as with P_2 and P_3 in the figure).

Notice, as in Fig. 2.1, that a directed graph may have separate "components" of vertices which are connected only among themselves; and some vertices, such as P_5 , may not be connected with any other vertex. Also, since $P_i \rightarrow P_i$ is not permitted in a directed graph, a vertex cannot be connected with itself by a single arc which does not pass through any other vertex.

In Fig. 2.2 we have drawn diagrams representing three more examples of directed graphs.

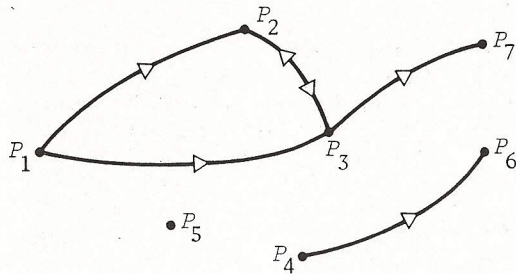


Figure 2.1

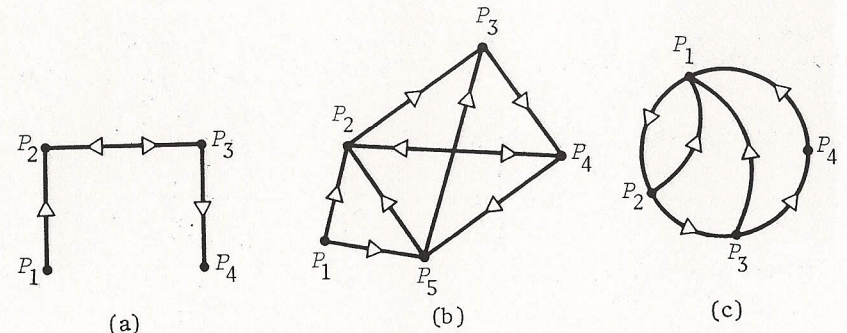


Figure 2.2

With a directed graph having n vertices, we may associate an $n \times n$ matrix $M = [m_{ij}]$, called the *vertex matrix* of the directed graph. Its elements are defined by

$$m_{ij} = \begin{cases} 1, & \text{if } P_i \rightarrow P_j \\ 0, & \text{otherwise} \end{cases}$$

for $i, j = 1, 2, \dots, n$. For the three directed graphs in Fig. 2.2, the corresponding vertex matrices are:

Figure 2.2(a):

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

Figure 2.2(b):

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

Figure 2.2(c):

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

By their definition, vertex matrices have the following two properties:

- (i) All entries are either 0 or 1.
- (ii) All diagonal entries are 0.

Conversely, any matrix with these two properties determines a unique directed graph having the matrix as its vertex matrix. For example, the matrix

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

determines the directed graph in Fig. 2.3.

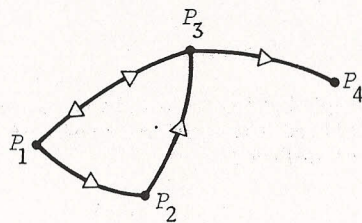


Figure 2.3

EXAMPLE 2.1 A certain family consists of a mother, father, daughter, and two sons. The family members have influence, or power, over each other in the following ways: the mother can influence the daughter and the oldest son; the father can influence the two sons; the daughter can influence the father; the oldest son can influence the youngest son; and the youngest son can influence the mother. We may model this family influence pattern with a directed graph whose vertices are the five family members. If family member *A* influences family member *B*, we write *A* → *B*. Figure 2.4 is the resulting directed graph, where we have used obvious letter designations for the five family members. The vertex matrix of this directed graph is

	<i>M</i>	<i>F</i>	<i>D</i>	<i>OS</i>	<i>YS</i>
<i>M</i>	0	0	1	1	0
<i>F</i>	0	0	0	1	1
<i>D</i>	0	1	0	0	0
<i>OS</i>	0	0	0	0	1
<i>YS</i>	1	0	0	0	0

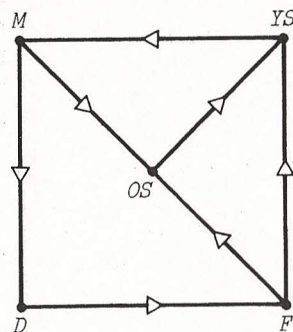


Figure 2.4

EXAMPLE 2.2 In chess, the knight moves in an "L"-shaped pattern about the chessboard. It may move horizontally two squares and then vertically one square, or it may move vertically two squares and then horizontally one square. For example, from the black square in Fig. 2.5, the knight may move to any of the eight shaded squares. Suppose the knight is restricted to the nine numbered squares in Fig. 2.6. If by *i* → *j* we mean that the knight may move from square *i* to square *j*, the directed graph in Figure 2.7 illustrates all possible moves that the knight may make among these nine squares. In Fig. 2.8 we have "unraveled" Fig. 2.7 to make the pattern of possible moves clearer.

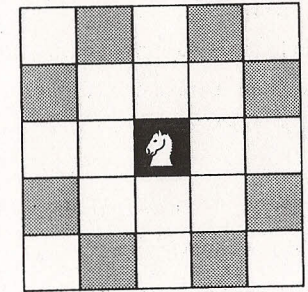


Figure 2.5

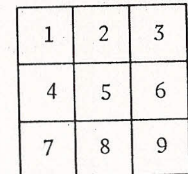


Figure 2.6

The vertex matrix of this directed graph is given by

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

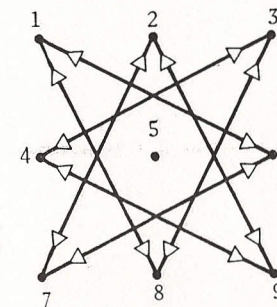


Figure 2.7

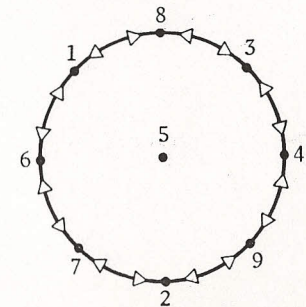


Figure 2.8

In Example 2.1, the father cannot directly influence the mother; i.e., $F \rightarrow M$ is not true. But he can influence the youngest son, who can then influence the mother. We write this as $F \rightarrow YS \rightarrow M$, and call it a 2-step connection from F to M . Analogously, we call $M \rightarrow D$ a 1-step connection; $F \rightarrow OS \rightarrow YS \rightarrow M$ a 3-step connection; and so forth. Let us now consider a technique for finding the number of all possible r -step connections ($r = 1, 2, \dots$) from one vertex P_i to another vertex P_j of an arbitrary directed graph. (This will include the case when P_i and P_j are the same vertex.) The number of 1-step connections from P_i to P_j is simply m_{ij} . That is, there is either zero or one 1-step connection from P_i to P_j depending on whether m_{ij} is zero or one. For the number of 2-step connections, we consider the square of the vertex matrix. If we let $m_{ij}^{(2)}$ be the (i, j) -th element of M^2 , we have

$$m_{ij}^{(2)} = m_{i1}m_{1j} + m_{i2}m_{2j} + \dots + m_{in}m_{nj}. \quad (2.1)$$

Now, if $m_{i1} = m_{1j} = 1$, there is a 2-step connection $P_i \rightarrow P_1 \rightarrow P_j$ from P_i to P_j . But if either m_{i1} or m_{1j} is zero, such a 2-step connection is not possible. Thus $P_i \rightarrow P_1 \rightarrow P_j$ is a 2-step connection if and only if $m_{i1}m_{1j} = 1$. Similarly, for any $k = 1, 2, \dots, n$, $P_i \rightarrow P_k \rightarrow P_j$ is a 2-step connection from P_i to P_j if and only if the term $m_{ik}m_{kj}$ on the righthand side of (2.1) is one; otherwise the term is zero.

Thus, the righthand side of (2.1) is the total number of two 2-step connections from P_i to P_j .

This same argument will work for finding the number of 3-, 4-, ... step connections from P_i to P_j . The general result is the following:

THEOREM 2.1 Let M be the vertex matrix of a directed graph and let $m_{ij}^{(r)}$ be the (i, j) -th element of M^r . Then $m_{ij}^{(r)}$ is equal to the number of r -step connections from P_i to P_j .

EXAMPLE 2.3 Figure 2.9 is the route map of a small airline which services the four cities P_1, P_2, P_3 , and P_4 . As a directed graph, its vertex matrix is

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

We have that

$$M^2 = \begin{bmatrix} 2 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 0 \\ 2 & 0 & 1 & 1 \end{bmatrix}$$

and

$$M^3 = \begin{bmatrix} 1 & 3 & 3 & 1 \\ 2 & 2 & 3 & 1 \\ 4 & 0 & 2 & 2 \\ 1 & 3 & 3 & 1 \end{bmatrix}$$

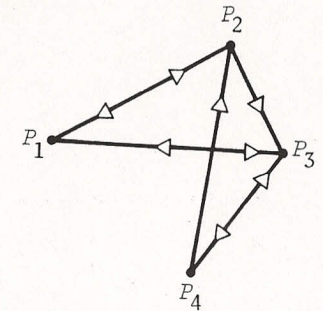


Figure 2.9

If we are interested in connections from city P_4 to city P_3 , we may use Theorem 2.1 to find their number. Since $m_{43} = 1$, there is one 1-step connection; since $m_{43}^{(2)} = 1$, there is one 2-step connection; and since $m_{43}^{(3)} = 3$, there are three 3-step connections. To verify this, from Fig. 2.9 we find:

- 1-step connections from P_4 to P_3 : $P_4 \rightarrow P_3$
- 2-step connections from P_4 to P_3 : $P_4 \rightarrow P_2 \rightarrow P_3$
- 3-step connections from P_4 to P_3 : $P_4 \rightarrow P_3 \rightarrow P_4 \rightarrow P_3$
 $P_4 \rightarrow P_2 \rightarrow P_1 \rightarrow P_3$
 $P_4 \rightarrow P_3 \rightarrow P_1 \rightarrow P_3$

CLIQUEES

Let us introduce the following definition:

DEFINITION 2.1 A subset of a directed graph is called a **clique** if it satisfies the following three conditions:

- (i) The subset contains at least three vertices.
- (ii) For each pair of vertices P_i and P_j in the subset, both $P_i \rightarrow P_j$ and $P_j \rightarrow P_i$ are true.
- (iii) The subset is as large as possible; i.e., it is not possible to add another vertex to the subset and still satisfy condition (ii).

This definition suggests that cliques are maximal subsets which are in perfect "communication" with each other. For example, if the vertices represent cities, and $P_i \rightarrow P_j$ means that there is a direct airline flight from city P_i to city P_j , then there is a direct flight between any two cities within a clique in either direction.

EXAMPLE 2.4 The directed graph illustrated in Fig. 2.10 (which might represent the route map of an airline) has two cliques:

$\{P_1, P_2, P_3, P_4\}$ and $\{P_3, P_4, P_6\}$.

This example shows that a directed graph may contain several cliques, and that a vertex may simultaneously belong to more than one clique.

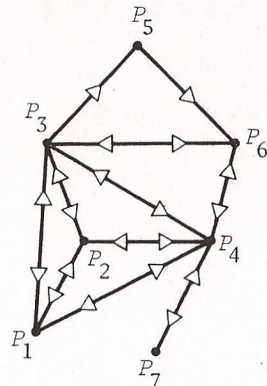


Figure 2.10

For simple directed graphs, cliques may be found by inspection. But for large directed graphs, it would be desirable to have a systematic procedure for detecting cliques. We shall now discuss a theorem which identifies those vertices which belong to cliques. First, we define a matrix $S = [s_{ij}]$ related to a given directed graph as follows:

$$s_{ij} = \begin{cases} 1, & \text{if } P_i \leftrightarrow P_j \\ 0, & \text{otherwise.} \end{cases}$$

The matrix S determines a directed graph which is the same as the given directed graph with the exception that the directed edges with only one arrow are deleted. For example, if the original directed graph is given by Fig. 2.11(a),

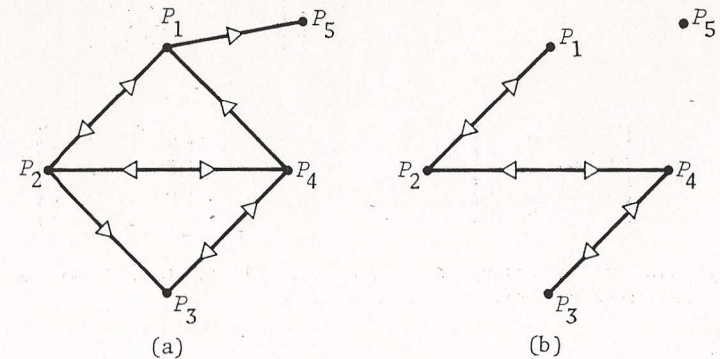


Figure 2.11

the directed graph which has S as its vertex matrix is given in Fig. 2.11(b). Alternately, S may be obtained from the vertex matrix M of the original directed graph by setting $s_{ij} = 1$ if $m_{ij} = m_{ji} = 1$, and otherwise setting $s_{ij} = 0$.

Using this matrix S , the theorem referred to is

THEOREM 2.2 Let $s_{ij}^{(3)}$ be the (i, j) -th element of S^3 . Then a vertex P_i belongs to some clique if and only if $s_{ii}^{(3)} \neq 0$.

PROOF If $s_{ii}^{(3)} \neq 0$, then there is at least one 3-step connection from P_i to itself in the modified directed graph determined by S . Suppose it is $P_i \rightarrow P_j \rightarrow P_k \rightarrow P_i$. In the modified directed graph, all directed relations are two-way, so that we also have $P_i \leftrightarrow P_j \leftrightarrow P_k \leftrightarrow P_i$. But this means either $\{P_i, P_j, P_k\}$ is a clique, or is a subset of a clique. In either case, P_i must belong to some clique. The converse statement that if P_i belongs to a clique then $s_{ii}^{(3)} \neq 0$ follows in a similar manner. \square

EXAMPLE 2.5 Suppose a directed graph has as its vertex matrix

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

Then

$$S = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad S^3 = \begin{bmatrix} 0 & 3 & 0 & 2 \\ 3 & 0 & 2 & 0 \\ 0 & 2 & 0 & 1 \\ 2 & 0 & 1 & 0 \end{bmatrix}$$

Since all diagonal entries of S^3 are zero, it follows from Theorem 2.2 that the directed graph has no cliques.

EXAMPLE 2.6 Suppose a directed graph has as its vertex matrix

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

Then

$$S = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad S^3 = \begin{bmatrix} 2 & 4 & 0 & 4 & 3 \\ 4 & 2 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 4 & 3 & 0 & 2 & 1 \\ 3 & 1 & 0 & 1 & 0 \end{bmatrix}$$

The nonzero diagonal entries of S^3 are $s_{11}^{(3)}$, $s_{22}^{(3)}$, and $s_{44}^{(3)}$. Consequently, in the given directed graph, P_1 , P_2 , and P_4 belong to cliques. Since a clique must contain at least three vertices, the directed graph has only one clique; namely, $\{P_1, P_2, P_4\}$.

DOMINANCE DIRECTED GRAPHS

In many groups of individuals or animals, there is a definite "pecking order" or dominance relation between any two members of the group. That is, given any two individuals A and B , either A dominates B or B dominates A , but not both. In terms of a directed graph

in which $P_i \rightarrow P_j$ means P_i dominates P_j , this means that for all distinct pairs, either $P_i \rightarrow P_j$ or $P_j \rightarrow P_i$, but not both. In general, we have the following definition:

DEFINITION 2.2 A dominance directed graph is a directed graph such that for any distinct pair of vertices P_i and P_j , either $P_i \rightarrow P_j$ or $P_j \rightarrow P_i$, but not both.

An example which provides a directed graph satisfying this definition is a league of n sports teams which play each other exactly one time, as in one round of a round-robin tournament in which no ties are allowed. If $P_i \rightarrow P_j$ means that team P_i beat team P_j in their single match, it is easily seen that Definition 2.2 is satisfied. For this reason, dominance directed graphs are sometimes called *tournaments*.

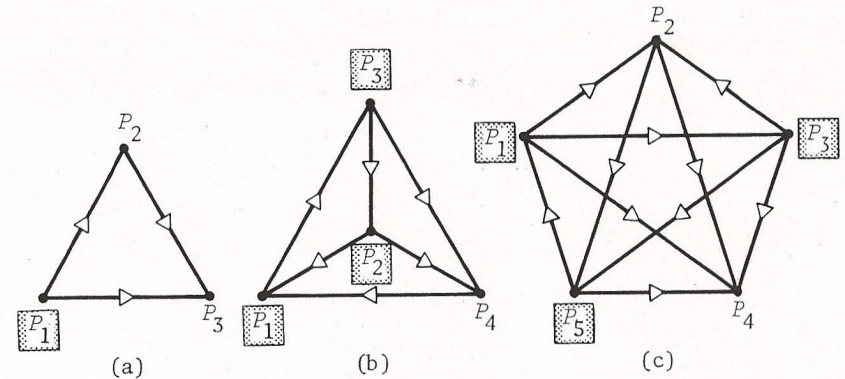


Figure 2.12

In Fig. 2.12 there are illustrated some dominance directed graphs with three, four, and five vertices, respectively. In these three graphs, the shaded vertices have the following interesting property: from each one there is either a 1-step or a 2-step connection to any other vertex in its graph. In a sense, these vertices are more "powerful" than the vertices which do not have this property. In a sports tournament, such a vertex corresponds to a team A such that if B is any other team, either A beat B , or beat a team which beat B . We shall now state and prove a theorem that guarantees that any dominance directed graph has at least one vertex with this property.

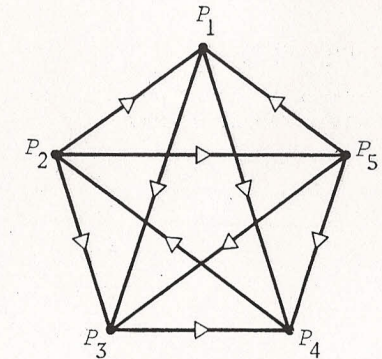
THEOREM 2.3 *In any dominance directed graph there is at least one vertex from which there is a 1-step or 2-step connection to any other vertex.*

PROOF. Consider a vertex (there may be several) with the largest total number of 1-step and 2-step connections to other vertices in the graph. By renumbering the vertices, we may assume that P_1 is such a vertex. Suppose there is some vertex P_i such that there is no 1-step or 2-step connection from P_1 to P_i . Then, in particular, $P_1 \rightarrow P_i$ is not true, so that, by Definition 2.2, it must be that $P_i \rightarrow P_1$. Next, let P_k be any vertex such that $P_1 \rightarrow P_k$ is true. Then we cannot have $P_k \rightarrow P_i$, since then $P_1 \rightarrow P_k \rightarrow P_i$ would be a 2-step connection from P_1 to P_i . Thus, it must be that $P_i \rightarrow P_k$. That is, P_i has 1-step connections to all the vertices to which P_1 has 1-step connections. The vertex P_i must then also have 2-step connections to all the vertices to which P_1 has 2-step connections. But since, in addition, we have that $P_i \rightarrow P_1$, this means that P_i has more 1-step and 2-step connections to other vertices than does P_1 . However, this is a contradiction to the way P_1 was chosen. Hence, there can be no vertex P_i to which P_1 has no 1-step or 2-step connection. \square

This proof shows that a vertex with the largest total number of 1-step and 2-step connections to other vertices has the property stated in the theorem. There is a simple way of finding such vertices by using the vertex matrix M and its square M^2 . The sum of the entries in the i -th row of M is the total number of 1-step connections from P_i to other vertices, and the sum of the entries of the i -th row of M^2 is the total number of 2-step connections from P_i to other vertices. Consequently, the sum of the entries of the i -th row of the matrix $A = M + M^2$ is the total number of 1-step and 2-step connections from P_i to other vertices. In other words, a row of $A = M + M^2$ with the largest row sum identifies a vertex having the property stated in Theorem 2.3.



EXAMPLE 2.7 Suppose five baseball teams play each other exactly once, and the results are as indicated in the dominance directed graph of Fig. 2.13. The vertex matrix of the graph is



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

And so

Figure 2.13

$$A = M + M^2 = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 3 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 2 & 0 \\ 2 & 0 & 3 & 3 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 2 & 3 & 0 \end{bmatrix}$$

The row sums of A are:

- 1st row sum = 4
- 2nd row sum = 9
- 3rd row sum = 2
- 4th row sum = 4
- 5th row sum = 7.

Since the second row has the largest row sum, the vertex P_2 must have a 1-step or 2-step connection to any other vertex. This is easily verified from Fig. 2.13.

We have informally suggested that a vertex with the largest number of 1-step and 2-step connections to other vertices is a "powerful" vertex. Let us formalize this concept with the following definition:

DEFINITION 2.3 *The power of a vertex of a dominance directed graph is the total number of 1-step and 2-step connections from it to other vertices. Alternately, the power of a vertex P_i is the sum of the entries of the i -th row of the matrix $A = M + M^2$, where M is the vertex matrix of the directed graph.*

EXAMPLE 2.8 Let us rank the five baseball teams in Example 2.7 according to their powers. From the calculations for the row sums in that example we have:

- Power of team $P_1 = 4$
- Power of team $P_2 = 9$
- Power of team $P_3 = 2$
- Power of team $P_4 = 4$
- Power of team $P_5 = 7$

Hence, the ranking of the teams according to their powers would be: first, P_2 ; second, P_5 ; third, P_1 and P_4 (tie); fifth, P_3 .

EXERCISES

2.1 Construct the vertex matrix for each of the directed graphs illustrated in Fig. 2.14.

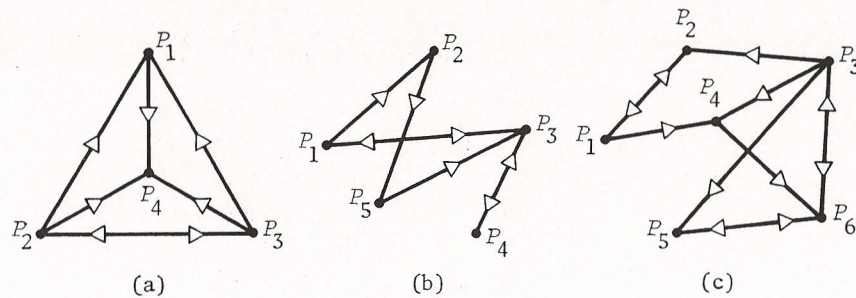


Figure 2.14

2.2 Draw a diagram of the directed graph corresponding to each of the following vertex matrices.

(a) $\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}$

2.3 Let M be the following vertex matrix of a directed graph:

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

- (a) Draw a diagram of the directed graph.
- (b) Use Theorem 2.1 to find the number of 1-, 2-, and 3-step connections from the vertex P_1 to the vertex P_2 . Verify your answer by listing the various connections as in Example 2.3.
- (c) Repeat part (b) for the 1-, 2-, and 3-step connections from P_1 to P_4 .

2.4 By inspection, locate all cliques in each of the directed graphs illustrated in Fig. 2.15.

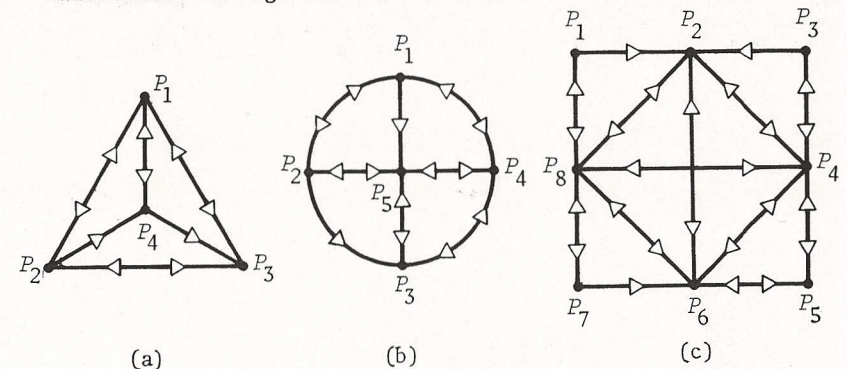


Figure 2.15

2.5 For each of the following vertex matrices, use Theorem 2.2 to find any cliques in the corresponding directed graphs.

(a) $\begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}$

(b) $\begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix}$

2.6 For the dominance directed graph illustrated in Fig. 2.16, construct the vertex matrix and use Definition 2.3 to find the power of each vertex.

2.7 Five baseball teams play each other one time with the following results:

- A beats B, C, D
- B beats C, E
- C beats D, E
- D beats B
- E beats A, D.

Rank the five baseball teams in accordance with the powers of the vertices they correspond to in the dominance directed graph representing the outcomes of the games.

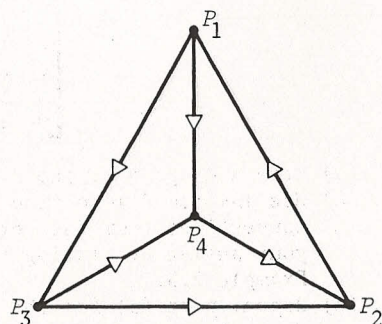


Figure 2.16

Theory of Games

3

A general game in which two competing players choose separate strategies to reach opposing objectives is discussed. Using matrix techniques, the optimal strategy of each player is found for a few special cases.

PREREQUISITES:

Matrix multiplication

INTRODUCTION

To introduce the basic concepts in the theory of games, let us consider the following carnival-type game which two people agree to play. We shall call the participants of the game player R and player C . Each player has a stationary wheel with a movable pointer on it as in Fig. 3.1. For reasons which will become clear, we shall call player R 's wheel the row-wheel, and player C 's wheel the column-wheel. The row-wheel is divided into three sectors numbered 1, 2, and 3, and the column wheel is divided into four sectors numbered 1, 2, 3, and 4. The fractions of the area occupied by the various sectors are indicated in the figure. To play the game, each player spins the pointer of his wheel and lets it come to rest at random. The number of the sector in which each pointer comes to

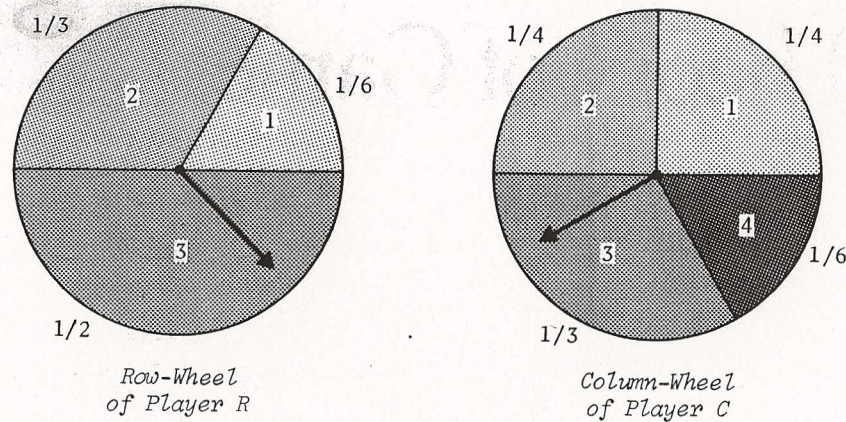


Figure 3.1

rest is called the *move* of that player. Thus, player *R* has three possible moves, and player *C* has four possible moves. Depending on the move each player makes, player *C* then makes a payment of money to player *R* according to the following table:

		Player <i>C</i> 's move			
		1	2	3	4
Player <i>R</i> 's move	1	\$3	\$5	-\$2	-\$1
	2	-\$2	\$4	-\$3	-\$4
	3	\$6	-\$5	\$0	\$3

Table 3.1

For example, if the row-wheel pointer comes to rest in sector 1 (player *R* makes move 1), and the column-wheel pointer comes to rest in sector 2 (player *C* makes move 2), then player *C* must pay player *R* the sum of \$5. Some of the entries in this table are negative, indicating that player *C* makes a negative payment to player *R*. By this we mean that player *R* makes a positive payment to player *C*. For example, if the row-wheel shows 2 and the column wheel shows 4, then player *R* pays player *C* the sum of \$4, since the corresponding entry in the table is -\$4. In this way, the positive entries of the table are the gains of player *R* and the losses of player *C*, and the negative entries are the gains of player *C* and the losses of player *R*.

In this game, the players have no control over their moves; each move is determined by chance. However, if each player can decide whether or not he wants to play, then each would want to know how much he can expect to win or lose over the long term if he chooses to play. We discuss this question in the first part of the next section. Later in the section we consider a more complicated situation in which the players can exercise some control over their moves by varying the sectors of their wheels.

TWO-PERSON ZERO-SUM MATRIX GAMES

The game described above is an example of a two-person zero-sum matrix game. The term "zero-sum" means that in each play of the game the positive gain of one player is equal to the negative gain (loss) of the other player. That is, the sum of the two algebraic gains is zero. The term "matrix game" is used to describe a two-person game in which each player has only a finite number of moves, so that all possible outcomes of each play, and the corresponding gains of the players, may be displayed in tabular or matrix form, as in Table 3.1.

In a general game of this type, let player *R* have *m* possible moves and let player *C* have *n* possible moves. In a play of the game, each player makes one of his possible moves, and then a *payoff* is made from player *C* to player *R* depending on the moves. For *i* = 1, 2, ..., *m* and *j* = 1, 2, ..., *n*, let us set

a_{ij} = payoff that player *C* makes to player *R* if player *R* makes move *i* and player *C* makes move *j*.

This payoff need not be money; it may be any type of commodity to which we can attach a numerical value. As before, if an entry a_{ij} is negative, we mean that player *C* receives a payoff of $|a_{ij}|$ from player *R*. We arrange these *mn* possible payoffs in the form of an *m* × *n* matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

which we shall call the *payoff matrix* of the game.

Each player is to make his moves on a probabilistic basis. For example, for the game discussed in the introduction, the proportion of the area of a sector to the area of the wheel would be the probability that the player makes the move corresponding to that sector.

Thus, from Fig. 3.1, we see that player R would make move 2 with probability $1/3$, and player C would make move 2 with probability $1/4$. In the general case, let us set

p_i = probability that player R makes move i ($i = 1, 2, \dots, m$),

and

q_j = probability that player C makes move j ($j = 1, 2, \dots, n$).

By their definitions, we have that

$$p_1 + p_2 + \dots + p_m = 1$$

and

$$q_1 + q_2 + \dots + q_n = 1.$$

With these probabilities, p_i and q_j , we form two vectors

$$\mathbf{p} = [p_1 \ p_2 \ \dots \ p_m] \quad \text{and} \quad \mathbf{q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix}.$$

We call the row vector \mathbf{p} the *strategy of player R*, and call the column vector \mathbf{q} the *strategy of player C*. For example, from Fig. 3.1 we have

$$\mathbf{p} = [1/6 \ 1/3 \ 1/2] \quad \text{and} \quad \mathbf{q} = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/3 \\ 1/6 \end{bmatrix}$$

for the game described in the introduction.

From the theory of probability, if the probability that player R makes move i is p_i , and independently the probability that player C makes move j is q_j , then $p_i q_j$ is the probability that for any one play of the game player R makes move i and player C makes move j . The payoff to player R for such a pair of moves is a_{ij} . Thus $p_i q_j$ is also the probability that for any one play of the game the payoff to player R is entry a_{ij} . If we multiply each possible payoff by its

corresponding probability and sum over all possible payoffs, we obtain the expression

$$a_{11}p_1q_1 + a_{12}p_1q_2 + a_{21}p_2q_1 + \dots + a_{mn}p_mq_n. \quad (3.1)$$

Equation (3.1) is a weighted average of the payoffs to player R , with each payoff being weighted according to the probability of its occurrence. In the theory of probability this weighted average is called the *expected payoff* to player R . It can be shown that if the game is played many times, the long term average payoff per play to player R is given by this expression. We shall denote this expected payoff by $E(\mathbf{p}, \mathbf{q})$ to emphasize the fact that it depends on the strategies of the two players. From the definition of the payoff matrix A and the strategies \mathbf{p} and \mathbf{q} , it may be verified that we may express the expected payoff in matrix notation as

$$E(\mathbf{p}, \mathbf{q}) = [p_1 \ p_2 \ \dots \ p_m] \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix} = \mathbf{p}A\mathbf{q}. \quad (3.2)$$

Observe that since $E(\mathbf{p}, \mathbf{q})$ is the expected payoff to player R , then $-E(\mathbf{p}, \mathbf{q})$ is the expected payoff to player C .

EXAMPLE 3.1 For the game described in the introduction, we have

$$E(\mathbf{p}, \mathbf{q}) = \mathbf{p}A\mathbf{q} = [1/6 \ 1/3 \ 1/2] \begin{bmatrix} 3 & 5 & -2 & -1 \\ -2 & 4 & -3 & -4 \\ 6 & -5 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/4 \\ 1/3 \\ 1/6 \end{bmatrix} = 13/72 = .1806\dots$$

Thus, in the long run player R can expect to receive an average of about 18¢ from player C in each play of the game.

So far we have been discussing the situation in which each player has a predetermined strategy. We shall now consider the more difficult situation in which both players may change their strategies independently. For example, in the game described in the introduction, we would allow both players to alter the areas of the sectors of their wheels and thereby control the probabilities of their respective moves. This qualitatively changes the nature of the problem and puts us firmly into the field of true game theory. It is understood that neither player knows the strategy the other

will choose. It will also be assumed that each player will make the best possible choice of strategy, and that the other player knows this. Thus player R attempts to choose a strategy p such that $E(p, q)$ is as large as possible for the best strategy q that player C can choose; and similarly player C attempts to choose a strategy q such that $E(p, q)$ is as small as possible for the best strategy p that player R can choose. First, to see that such choices are possible, let us state the following theorem, called the *Fundamental Theorem of 2-person Zero-sum Games*:

THEOREM 3.1 *There exist strategies p^* and q^* such that*

$$E(p^*, q) \geq E(p^*, q^*) \geq E(p, q^*) \quad (3.3)$$

for all strategies p and q .

(The proof of this theorem involves ideas from the theory of Linear Programming, and we shall not give it here. In the next two sections we shall prove the theorem for two special cases.)

This result guarantees that each player has a best possible strategy. To see why, let p^* and q^* be the strategies stated in the theorem, and let $v = E(p^*, q^*)$. The lefthand inequality of Eq. (3.3) then reads

$$E(p^*, q) \geq v, \quad \text{for all strategies } q.$$

This means that if player R chooses the strategy p^* , then no matter what strategy q player C chooses, the expected payoff to player R will never be below v . Furthermore, the expected payoff v is the best possible. For suppose there is some strategy p^{**} that player R can choose such that

$$E(p^{**}, q) > v, \quad \text{for all strategies } q.$$

Then, in particular,

$$E(p^{**}, q^*) > v.$$

But this is a contradiction to the righthand inequality of Eq. (3.3) which requires that $E(p^{**}, q^*) \leq v$. Consequently, the best player R can do is prevent his expected payoff from falling below the value v .

Similarly, if player C chooses the strategy q^* , then from the righthand inequality of Eq. (3.3) we have

$$v \geq E(p, q^*), \quad \text{for all strategies } p.$$

That is, the expected payoff to player R never exceeds the value v if player C chooses the strategy q^* , no matter what strategy player R chooses. This bound similarly cannot be lowered, since player R may always choose the strategy p^* to force his expected payoff to be at least the value v .

On the basis of this discussion, we make the following definitions:

DEFINITION 3.1 *If p^* and q^* are strategies such that*

$$E(p^*, q) \geq E(p^*, q^*) \geq E(p, q^*) \quad (3.4)$$

for all strategies p and q , then

- (i) p^* is called an optimal strategy for player R
- (ii) q^* is called an optimal strategy for player C ,
- (iii) $v = E(p^*, q^*)$ is called the value of the game.

The wording in this definition suggests that optimal strategies are not necessarily unique. This is indeed the case, and in Exercise 3.2 we ask the reader to show this. However, it can be proved that any two sets of optimal strategies always result in the same value v of the game. That is, if p^* , q^* and p^{**} , q^{**} are optimal strategies, then

$$E(p^*, q^*) = E(p^{**}, q^{**}). \quad (3.5)$$

The value of a game is thus the expected payoff to player R when both players choose any possible optimal strategies.

To find optimal strategies, we must find vectors p^* and q^* which satisfy condition (3.4). This is generally done by using Linear Programming techniques. (See Chapter 13 for an introduction to this field.) In the next two sections we discuss special cases for which optimal strategies may be found by more elementary techniques.

STRICTLY DETERMINED GAMES

Let us introduce the following definition:

DEFINITION 3.2 *An entry a_{rs} in a payoff matrix A is called a saddle point if*

(i) a_{rs} is the smallest entry in its row, and
 (ii) a_{rs} is the largest entry in its column.
 A game whose payoff matrix has a saddle point is called strictly determined.

For example, the shaded element in each of the following payoff matrices is a saddle point:

$$\begin{bmatrix} 3 & 1 \\ -4 & 0 \end{bmatrix} \quad \begin{bmatrix} 30 & -50 & -5 \\ 60 & 90 & 75 \\ -10 & 60 & -30 \end{bmatrix} \quad \begin{bmatrix} 0 & -3 & 5 & -9 \\ 15 & -8 & -2 & 10 \\ 7 & 10 & 6 & 9 \\ 6 & 11 & -3 & 2 \end{bmatrix}$$

If a matrix has a saddle point a_{rs} , it turns out that the following strategies are optimal strategies for the two players:

$$p^* = [0 \ 0 \ \dots \ 1 \ \dots \ 0], \quad q^* = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

↖ rth entry
← sth entry.

That is, an optimal strategy for player R is to always make the r th move, and an optimal strategy for player C is to always make the s th move. Such strategies for which only one move is possible are called *pure strategies*. Strategies for which more than one move is possible are called *mixed strategies*. To show that the above pure strategies are optimal, the reader may verify the following three equations (see Exercise 3.6):

$$E(p^*, q^*) = p^* A q^* = a_{rs}, \tag{3.6}$$

$$E(p^*, q) = p^* A q \geq a_{rs} \text{ for any strategy } q, \tag{3.7}$$

$$E(p, q^*) = p A q^* \leq a_{rs} \text{ for any strategy } p. \tag{3.8}$$

Together, these three equations imply that

$$E(p^*, q) \geq E(p^*, q^*) \geq E(p, q^*)$$

for all strategies p and q . Since this is exactly condition (3.4), it follows that p^* and q^* are optimal strategies.

From Eq. (3.6), the value of a strictly determined game is simply the numerical value of a saddle point a_{rs} . It is possible for a

payoff matrix to have several saddle points, but then the uniqueness of the value of a game guarantees that the numerical values of all saddle points are the same.

EXAMPLE 3.2 Two competing television networks, R and C , are scheduling 1-hour programs in the same time period. Network R can schedule one of three possible programs, and network C can schedule one of four possible programs. Neither network knows which program the other will schedule. Both networks ask the same outside polling agency to give them an estimate of how all possible pairings of the programs will divide the viewing audience. The agency gives them each the following table, whose (i, j) th entry is the percentage of the viewing audience that will watch network R if network R 's program i is paired against network C 's program j .

		Network C 's Program			
		1	2	3	4
Network R 's Program	1	60	20	30	55
	2	50	75	45	60
	3	70	45	35	30

What program should each network schedule in order to maximize its viewing audience?

SOLUTION Let us subtract 50 from each entry in the above table to construct the following matrix:

$$\begin{bmatrix} 10 & -30 & -20 & 5 \\ 0 & 25 & -5 & 10 \\ 20 & -5 & -15 & -20 \end{bmatrix}$$

This is the payoff matrix of the two-person zero-sum game in which each network is considered to start with 50% of the audience, and the (i, j) th entry of the matrix is the percentage of the viewing audience that network C loses to network R if programs i and j are paired against each other. It is easily seen that the entry

$$a_{23} = -5$$

is a saddle point of the payoff matrix. Hence, the optimal strategy of network R is to schedule program 2 and the optimal strategy of network C is to schedule program 3. This will result in network R receiving 45% of the audience and network C receiving 55% of the audience.

2 x 2 MATRIX GAMES

Another case in which the optimal strategies may be found by elementary means is when each player has only two possible moves. In this case, the payoff matrix is a 2×2 matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

If the game is strictly determined, at least one of the four entries of A is a saddle point, and the technique of the previous section can then be applied to determine optimal strategies for the two players. If the game is not strictly determined, we proceed as described below. First, we compute the expected payoff for arbitrary strategies \mathbf{p} and \mathbf{q} :

$$\begin{aligned} E(\mathbf{p}, \mathbf{q}) &= \mathbf{p}A\mathbf{q} = \begin{bmatrix} p_1 & p_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \\ &= a_{11}p_1q_1 + a_{12}p_1q_2 + a_{21}p_2q_1 + a_{22}p_2q_2. \end{aligned} \quad (3.9)$$

Since

$$p_1 + p_2 = 1 \quad \text{and} \quad q_1 + q_2 = 1 \quad (3.10)$$

we may substitute $p_2 = 1 - p_1$ and $q_2 = 1 - q_1$ into (3.9) to obtain

$$E(\mathbf{p}, \mathbf{q}) = a_{11}p_1q_1 + a_{12}p_1(1 - q_1) + a_{21}(1 - p_1)q_1 + a_{22}(1 - p_1)(1 - q_1). \quad (3.11)$$

If we rearrange the terms in Eq. (3.11), we may write

$$E(\mathbf{p}, \mathbf{q}) = [(a_{11} + a_{22} - a_{12} - a_{21})p_1 - (a_{22} - a_{21})]q_1 + (a_{12} - a_{22})p_1 + a_{22}. \quad (3.12)$$

By examining the coefficient of the q_1 term in (3.12), we see that if we set

$$p_1 = p_1^* = \frac{a_{22} - a_{21}}{a_{11} + a_{22} - a_{12} - a_{21}}, \quad (3.13)$$

then that coefficient is zero, and Eq. (3.12) reduces to

$$E(\mathbf{p}^*, \mathbf{q}) = \frac{a_{11}a_{22} - a_{12}a_{21}}{a_{11} + a_{22} - a_{12} - a_{21}}. \quad (3.14)$$

Eq. (3.14) is independent of \mathbf{q} . That is, if player R chooses the strategy determined by (3.13), player C cannot change the expected payoff by varying his strategy.

In a similar manner, it may be verified that if player C chooses the strategy determined by

$$q_1 = q_1^* = \frac{a_{22} - a_{12}}{a_{11} + a_{22} - a_{12} - a_{21}}, \quad (3.15)$$

then substitution in Eq. (3.12) gives

$$E(\mathbf{p}, \mathbf{q}^*) = \frac{a_{11}a_{22} - a_{12}a_{21}}{a_{11} + a_{22} - a_{12} - a_{21}}. \quad (3.16)$$

Equations (3.14) and (3.16) show that

$$E(\mathbf{p}^*, \mathbf{q}) = E(\mathbf{p}^*, \mathbf{q}^*) = E(\mathbf{p}, \mathbf{q}^*) \quad (3.17)$$

for all strategies \mathbf{p} and \mathbf{q} . Thus, Definition 3.1 is satisfied, so that the strategies determined by Eqs. (3.13) and (3.15) are optimal strategies for players R and C , respectively. Let us summarize this in the following theorem, where we have computed the values of p_2^* and q_2^* from Eq. (3.10).

THEOREM 3.2 For a 2×2 game which is not strictly determined, optimal strategies for players R and C are

$$\mathbf{p}^* = \left[\frac{a_{22} - a_{21}}{a_{11} + a_{22} - a_{12} - a_{21}} \quad \frac{a_{11} - a_{12}}{a_{11} + a_{22} - a_{12} - a_{21}} \right]$$

and

$$q^* = \begin{bmatrix} \frac{a_{22} - a_{12}}{a_{11} + a_{22} - a_{12} - a_{21}} \\ \frac{a_{11} - a_{21}}{a_{11} + a_{22} - a_{12} - a_{21}} \end{bmatrix}$$

The value of the game is

$$v = \frac{a_{11}a_{22} - a_{12}a_{21}}{a_{11} + a_{22} - a_{12} - a_{21}}$$

In order to be complete, we must show that the entries in the vectors p^* and q^* are numbers between 0 and 1. In Exercise 3.7 we ask the reader to show that this is the case as long as the game is not strictly determined.

Equation (3.17) is interesting in that it implies that either player may force the expected payoff to be the value of the game by choosing his optimal strategy, regardless of which strategy the other player chooses. This is not true, in general, for games in which either player has more than two moves.

EXAMPLE 3.3 The Federal government desires to inoculate its citizens against a certain flu virus. The virus has two strains, and it is not known in what proportions the two strains occur in the virus population. Two vaccines have been developed with different effectivenesses against the two strains. Vaccine 1 is 85% effective against strain 1 and 70% effective against strain 2. Vaccine 2 is 60% effective against strain 1 and 90% effective against strain 2. What inoculation policy should the government adopt?

SOLUTION We may consider this a two-person game in which player R (the government) desires to make the payoff (the fraction of citizens resistant to the virus) as large as possible, and player C (the virus) desires to make the payoff as small as possible. The payoff matrix is

		Strain	
		1	2
Vaccine	1	.85	.70
	2	.60	.90

It can be seen that this matrix has no saddle points, so that Theorem 3.2 is applicable. Consequently,

$$p_1^* = \frac{a_{22} - a_{21}}{a_{11} + a_{22} - a_{12} - a_{21}} = \frac{.90 - .60}{.85 + .90 - .70 - .60} = \frac{.30}{.45} = \frac{2}{3}$$

$$p_2^* = 1 - p_1^* = 1 - \frac{2}{3} = \frac{1}{3}$$

$$q_1^* = \frac{a_{22} - a_{12}}{a_{11} + a_{22} - a_{12} - a_{21}} = \frac{.90 - .70}{.85 + .90 - .70 - .60} = \frac{.20}{.45} = \frac{4}{9}$$

$$q_2^* = 1 - q_1^* = 1 - \frac{4}{9} = \frac{5}{9}$$

$$v = \frac{a_{11}a_{22} - a_{12}a_{21}}{a_{11} + a_{22} - a_{12} - a_{21}} = \frac{(.85)(.90) - (.70)(.60)}{.85 + .90 - .70 - .60} = \frac{.345}{.45} = .7666\dots$$

Thus, the optimal strategy for the government is to inoculate 2/3 of the citizens with vaccine 1 and 1/3 of the citizens with vaccine 2. This will guarantee that 76.7% of the citizens will be resistant to a virus attack regardless of the distribution of the two strains.

On the other hand, a virus distribution of 4/9 of strain 1 and 5/9 of strain 2 will result in the same 76.7% of resistant citizens regardless of the inoculation strategy adopted by the government.

EXERCISES

3.1 Let a game have payoff matrix

$$A = \begin{bmatrix} -4 & 6 & -4 & 1 \\ 5 & -7 & 3 & 8 \\ -8 & 0 & 6 & -2 \end{bmatrix}$$

(a) If players R and C use strategies

$$p = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \text{ and } q = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix},$$

- respectively, what is the expected payoff of the game?
- (b) If player C keeps his strategy fixed as in part (a), what strategy should player R choose to maximize his expected payoff?
- (c) If player R keeps his strategy fixed as in part (a), what strategy should Player C choose to minimize the expected payoff to player R ?
- 3.2 Construct a simple example to show that optimal strategies defined in Definition 3.1 are not necessarily unique. For example, find a payoff matrix with several equal saddle points.
- 3.3 For the strictly determined games with the following payoff matrices, find optimal strategies for the two players and find the values of the games.

$$(a) \begin{bmatrix} 5 & 2 \\ 7 & 3 \end{bmatrix} \quad (b) \begin{bmatrix} -3 & -2 \\ 2 & 4 \\ -4 & 1 \end{bmatrix} \quad (c) \begin{bmatrix} 2 & -2 & 0 \\ -6 & 0 & -5 \\ 5 & 2 & 3 \end{bmatrix} \quad (d) \begin{bmatrix} -3 & 2 & -1 \\ -2 & -1 & 5 \\ -4 & 1 & 0 \\ -3 & 4 & 6 \end{bmatrix}$$

- 3.4 For the 2×2 games with the following payoff matrices, find optimal strategies for the two players and find the values of the games.

$$(a) \begin{bmatrix} 6 & 3 \\ -1 & 4 \end{bmatrix} \quad (b) \begin{bmatrix} 40 & 20 \\ -10 & 30 \end{bmatrix} \quad (c) \begin{bmatrix} 3 & 7 \\ -5 & 4 \end{bmatrix} \quad (d) \begin{bmatrix} 3 & 5 \\ 5 & 2 \end{bmatrix} \quad (e) \begin{bmatrix} 7 & -3 \\ -5 & -2 \end{bmatrix}$$

- 3.5 Player R has two playing cards: a black ace and a red four. Player C also has two cards: a black two and a red three. Each player secretly selects one of his cards. If both selected cards are the same color, player C pays player R the sum of the face values in dollars. If the cards are different colors, player R pays player C the sum of the face values. What are optimal strategies for both players and what is the value of the game?
- 3.6 Verify Equations (3.6), (3.7), and (3.8).
- 3.7 Show that the entries of the optimal strategies p^* and q^* given in Theorem 3.2 are numbers strictly between zero and one.

Markov Chains

4

A general model of a system which moves from state to state is described and applied to several concrete problems. It is shown that such systems tend to a steady-state configuration for large times.

PREREQUISITES:

Linear systems
Matrices
Intuitive understanding of limits

INTRODUCTION

Suppose a physical or mathematical system is such that at any moment it can occupy one of a finite number of states. For example, the weather in a certain city could be in one of three possible states: sunny, cloudy, or raining. Or an individual could be in one of four possible emotional states: happy, sad, angry, or apprehensive. Suppose such a system moves with time from one state to another. Let us construct a schedule of observation times and keep a record of the states of the system at these times. If we find that the transition from one state to another is not predetermined, but rather can only be specified in terms of certain probabilities which

depend on the previous history of the system, then the process is called a *stochastic process*. If, in addition, these transition probabilities depend only on the immediate history of the system, that is, if the state of the system at any observation is dependent only on its state at the immediately preceding observation, then the process is called a *Markov process* or a *Markov chain*.

Let the system under observation have k possible states, which we label as states $1, 2, \dots, k$. We define the following quantities to describe the transitions of the system from one state to another:

DEFINITION 4.1 The transition probability p_{ij} ($i, j = 1, 2, \dots, k$) is the probability that if the system is in state j at any one observation, it will be in state i at the next observation.

For example, if state 2 corresponds to a rainy day in Detroit, and state 3 corresponds to a cloudy day, then p_{32} is the probability that the weather in Detroit changes from rainy to cloudy in two successive days.

As probabilities, the numbers p_{ij} must all lie in the interval $[0, 1]$. Furthermore, for any fixed $j = 1, 2, \dots, k$, we must have:

$$p_{1j} + p_{2j} + \dots + p_{kj} = 1, \quad (4.1)$$

which expresses the fact that if the system is in state j at one observation, it will with certainty be in one of the k states at the next observation.

With these k^2 transition probabilities, we may form a $k \times k$ matrix $P = [p_{ij}]$, called the *transition matrix of the Markov process*. Equation (4.1) expresses the fact that the sum of the entries in each column of P is one. More generally, we have the following definition:

DEFINITION 4.2 A transition matrix $P = [p_{ij}]$ is any square matrix with nonnegative entries, all of whose column sums are one.

Thus, any Markov process determines a transition matrix. Transition matrices are also called *Markov matrices*, *probability matrices*, or *stochastic matrices*.

EXAMPLE 4.1 By reviewing its donation records, the alumni office of a college finds that 80% of its alumni who contribute to the annual fund one year will also contribute the next year, and 30% of those who do not contribute one year will contribute the next. This may be described in terms of a Markov process with two states: state 1 corresponds to an alumnus giving a donation in any one year, and state 2 corresponds to the alumnus not giving a donation in that year. The transition matrix of the process is

$$P = \begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix}.$$

Our objective in this chapter is to predict the state of a system described by a Markov process at future observation times. However, since the system is not deterministic, our predictions will be in terms of probabilities. To this end we introduce the following definition:

DEFINITION 4.3 A probability vector is a column vector with nonnegative entries whose sum is one.

We may then specify a future observation of a system described by a Markov process in terms of a probability vector as follows:

DEFINITION 4.4 The probability vectors $x^{(n)}$ for $n = 0, 1, \dots$ are said to be the *state vectors* of a Markov process if the i -th component $x_i^{(n)}$ of $x^{(n)}$ is the probability that the system is in the i -th state at the n -th observation.

In particular, the state vector $x^{(0)}$ is called the *initial state vector* of the Markov process. Using the next theorem, we shall see that the initial state vector determines all future state vectors.

THEOREM 4.1 If P is the transition matrix of a Markov process and $x^{(n)}$ is the state vector at the n -th observation, then $x^{(n+1)} = Px^{(n)}$.

The proof of this theorem involves ideas from probability theory and will not be given here. From this theorem it follows that

$$\begin{aligned} \mathbf{x}^{(1)} &= P\mathbf{x}^{(0)} \\ \mathbf{x}^{(2)} &= P\mathbf{x}^{(1)} = P^2\mathbf{x}^{(0)} \\ \mathbf{x}^{(3)} &= P\mathbf{x}^{(2)} = P^3\mathbf{x}^{(0)} \\ &\vdots \\ \mathbf{x}^{(n)} &= P\mathbf{x}^{(n-1)} = P^n\mathbf{x}^{(0)}. \end{aligned}$$

In this way, the initial state vector $\mathbf{x}^{(0)}$ and the transition matrix P determine $\mathbf{x}^{(n)}$ for $n=1, 2, \dots$.

EXAMPLE 4.1 (REVISITED) The transition matrix in Example 4.1 was

$$P = \begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix}.$$

Let us construct the probable future donation record of a new graduate who did not give a donation in the initial year after his graduation. In this way, the system is initially in state 2 with certainty, and so the initial state vector is

$$\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

From Theorem 4.1, we then have

$$\mathbf{x}^{(1)} = P\mathbf{x}^{(0)} = \begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} .3 \\ .7 \end{bmatrix}$$

$$\mathbf{x}^{(2)} = P\mathbf{x}^{(1)} = \begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix} \begin{bmatrix} .3 \\ .7 \end{bmatrix} = \begin{bmatrix} .45 \\ .55 \end{bmatrix}$$

$$\mathbf{x}^{(3)} = P\mathbf{x}^{(2)} = \begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix} \begin{bmatrix} .45 \\ .55 \end{bmatrix} = \begin{bmatrix} .525 \\ .475 \end{bmatrix}.$$

Thus, after three years, the alumnus can be expected to make a donation with probability .525. Beyond three years, we find the following state vectors (to three decimal places):

$$\mathbf{x}^{(4)} = \begin{bmatrix} .563 \\ .438 \end{bmatrix}, \quad \mathbf{x}^{(5)} = \begin{bmatrix} .581 \\ .419 \end{bmatrix}, \quad \mathbf{x}^{(6)} = \begin{bmatrix} .591 \\ .409 \end{bmatrix}, \quad \mathbf{x}^{(7)} = \begin{bmatrix} .595 \\ .405 \end{bmatrix},$$

$$\mathbf{x}^{(8)} = \begin{bmatrix} .598 \\ .402 \end{bmatrix}, \quad \mathbf{x}^{(9)} = \begin{bmatrix} .599 \\ .401 \end{bmatrix}, \quad \mathbf{x}^{(10)} = \begin{bmatrix} .599 \\ .401 \end{bmatrix}, \quad \mathbf{x}^{(11)} = \begin{bmatrix} .600 \\ .400 \end{bmatrix}.$$

For all n beyond eleven, we have $\mathbf{x}^{(n)} = \begin{bmatrix} .600 \\ .400 \end{bmatrix}$ to three decimal

places. In other words, the state vectors converge to a fixed vector as the number of observations increases. We shall discuss this further in the next section.

EXAMPLE 4.2 A car rental agency has three rental locations, which we label as locations 1, 2, and 3. A customer may rent a car from any of the three locations and return the car to any of the three locations. The manager finds that the customers return the cars to the various locations according to the following probabilities:

Rented from location

	1	2	3	
.8	.3	.2	1	Returned to location
.1	.2	.6	2	
.1	.5	.2	3	

This matrix is the transition matrix of the system considered as a Markov process. The i -th component of a state vector is the probability that a car is returned to the i -th location. In the table below, we list the state vectors $\mathbf{x}^{(n)}$ for $n=0, 1, \dots, 11$ when a car is initially rented from location 2.

$\mathbf{x}^{(n)}$ \ n	0	1	2	3	4	5	6	7	8	9	10	11
$x_1^{(n)}$	0	.300	.400	.477	.511	.533	.544	.550	.553	.555	.556	.557
$x_2^{(n)}$	1	.200	.370	.252	.261	.240	.238	.233	.232	.231	.230	.230
$x_3^{(n)}$	0	.500	.230	.271	.228	.227	.219	.217	.215	.214	.214	.213

For all values of n greater than eleven, all state vectors are equal to $x^{(11)}$ to three decimal places.

Two things should be observed in this example. First, it was not necessary to know how long a customer kept his car. That is, in a Markov process the time period between observations need not be regular. Second, the state vectors approach a fixed vector as n increases, just as in the first example.

EXAMPLE 4.3 A traffic officer is assigned to control the traffic at the eight intersections indicated in Fig. 4.1. He is instructed to remain at each intersection for an hour and then to either remain at that same intersection or move to a neighboring intersection. To avoid establishing a pattern, he is told to choose his new intersection on a random basis, with each possible choice equally likely. For example, if he is at intersection 5, his next intersection can be 2, 4, 5, or 8, each with probability $1/4$. Every day he starts at the location where he stopped the day before. As a Markov process, we may construct the following transition matrix:

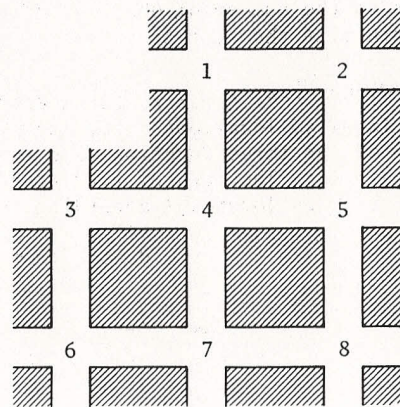


Figure 4.1

		Old Intersection								
		1	2	3	4	5	6	7	8	
New Intersection	1	1/3	1/3	0	1/5	0	0	0	0	1
	2	1/3	1/3	0	0	1/4	0	0	0	2
	3	0	0	1/3	1/5	0	1/3	0	0	3
	4	1/3	0	1/3	1/5	1/4	0	1/4	0	4
	5	0	1/3	0	1/5	1/4	0	0	1/3	5
	6	0	0	1/3	0	0	1/3	1/4	0	6
	7	0	0	0	1/5	0	1/3	1/4	1/3	7
	8	0	0	0	0	1/4	0	1/4	1/3	8

If the officer initially begins at intersection 5, his probable locations, hour by hour, are given by the following state vectors written in tabular form.

$x^{(n)}$ \ n	0	1	2	3	4	5	10	15	20	22
$x_1^{(n)}$	0	.000	.133	.116	.130	.123	.113	.109	.108	.107
$x_2^{(n)}$	0	.250	.146	.163	.140	.138	.115	.109	.108	.107
$x_3^{(n)}$	0	.000	.050	.039	.067	.073	.100	.106	.107	.107
$x_4^{(n)}$	0	.250	.113	.187	.162	.178	.178	.179	.179	.179
$x_5^{(n)}$	1	.250	.279	.190	.190	.168	.149	.144	.143	.143
$x_6^{(n)}$	0	.000	.000	.050	.056	.074	.099	.105	.107	.107
$x_7^{(n)}$	0	.000	.133	.104	.131	.125	.138	.141	.143	.143
$x_8^{(n)}$	0	.250	.146	.152	.124	.121	.108	.107	.107	.107

For all values of n greater than 22, all state vectors are equal to $x^{(22)}$ to three decimal places. So as with the first two examples, the state vectors approach a fixed vector as n increases.

LIMITING BEHAVIOR OF THE STATE VECTORS

In our three examples, we saw that the state vectors approached some fixed vector as the number of observations increased. We now ask whether the state vectors always approach a fixed vector in a Markov process. A simple example shows that this is not the case:

EXAMPLE 4.4

Let $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and $x^{(0)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Then, since $P^2 = I$ and $P^3 = P$, we have that

$$x^{(0)} = x^{(2)} = x^{(4)} = \dots = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and

$$x^{(1)} = x^{(3)} = x^{(5)} = \dots = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

This system oscillates indefinitely between the two state vectors $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, and so does not approach any fixed vector.

However, if we impose a mild condition on the transition matrix, we can show that a fixed limiting state vector is approached. This condition is described by the following definition:

DEFINITION 4.5 A transition matrix is regular if some integer power of it has all positive entries.

Thus, for a regular transition matrix P , there is some positive integer m such that all entries of P^m are positive. This is the case with the transition matrices of Examples 4.1 and 4.2 for $m=1$. In Example 4.3, it turns out that P^4 has all positive entries. Consequently, in all three examples the transition matrices are regular.

A Markov process which is governed by a regular transition matrix is called a *regular Markov process*. We shall see that every regular Markov process has a fixed state vector \mathbf{q} such that $P^n \mathbf{x}^{(0)}$ approaches \mathbf{q} as n increases for any choice of $\mathbf{x}^{(0)}$. This result is of major importance in the theory of Markov chains. It is based on the following theorem:

THEOREM 4.2 If P is a regular transition matrix, then as $n \rightarrow \infty$

$$P^n \rightarrow \begin{bmatrix} q_1 & q_1 & \dots & q_1 \\ q_2 & q_2 & \dots & q_2 \\ \vdots & \vdots & \ddots & \vdots \\ q_k & q_k & \dots & q_k \end{bmatrix}$$

where the q_i are positive numbers such that $q_1 + q_2 + \dots + q_k = 1$.

We shall not prove this theorem here. The interested reader is referred to a more specialized text, such as "Finite Markov Chains" by J. Kemeny and J. Snell (Springer-Verlag New York Inc., 1976).

Let us set

$$Q = \begin{bmatrix} q_1 & q_1 & \dots & q_1 \\ q_2 & q_2 & \dots & q_2 \\ \vdots & \vdots & \ddots & \vdots \\ q_k & q_k & \dots & q_k \end{bmatrix} \quad \text{and} \quad \mathbf{q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_k \end{bmatrix}.$$

Thus Q is a transition matrix, all of whose columns are equal to the probability vector \mathbf{q} . Q has the property that if \mathbf{x} is any probability vector, then

$$Q\mathbf{x} = \begin{bmatrix} q_1 & q_1 & \dots & q_1 \\ q_2 & q_2 & \dots & q_2 \\ \vdots & \vdots & \ddots & \vdots \\ q_k & q_k & \dots & q_k \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} q_1 x_1 + q_1 x_2 + \dots + q_1 x_k \\ q_2 x_1 + q_2 x_2 + \dots + q_2 x_k \\ \vdots \\ q_k x_1 + q_k x_2 + \dots + q_k x_k \end{bmatrix}$$

$$= (x_1 + x_2 + \dots + x_k) \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_k \end{bmatrix} = (1)\mathbf{q} = \mathbf{q}.$$

That is, Q transforms any probability vector \mathbf{x} into the fixed probability vector \mathbf{q} . From this, we can conclude the following:

THEOREM 4.3 If P is a regular transition matrix and \mathbf{x} is any probability vector, then as $n \rightarrow \infty$

$$P^n \mathbf{x} \rightarrow \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_k \end{bmatrix} = \mathbf{q},$$

where \mathbf{a} is a fixed probability vector independent of n all of whose entries are positive.

This follows since from Theorem 4.2 we have that $P^n \rightarrow Q$ as $n \rightarrow \infty$. This in turn implies that $P^n \mathbf{x} \rightarrow Q\mathbf{x} = \mathbf{a}$ as $n \rightarrow \infty$. \square

Thus, for a regular Markov process, the system eventually approaches a fixed state vector \mathbf{a} . The vector \mathbf{a} is called the *steady-state vector* of the regular Markov chain.

For systems with many states, usually the most efficient technique of computing the steady-state vector \mathbf{a} is simply to calculate $P^n \mathbf{x}$ for some large n . Our three examples illustrate this procedure. Each is a regular Markov process, so that convergence to a steady-state vector is assured. Another way of computing the steady-state vector is to make use of the following theorem:

THEOREM 4.4 *The steady-state vector \mathbf{a} of a regular transition matrix P is the unique probability vector which satisfies the equation $P\mathbf{a} = \mathbf{a}$.*

To see this, consider the matrix identity $PP^n = P^{n+1}$. By Theorem 4.1, both P^n and P^{n+1} approach Q as $n \rightarrow \infty$. Thus we have $PQ = Q$. Any one column of this matrix equation gives $P\mathbf{a} = \mathbf{a}$. To show that \mathbf{a} is the only probability vector which satisfies this equation, suppose \mathbf{r} is another probability vector such that $P\mathbf{r} = \mathbf{r}$. Then also $P^n \mathbf{r} = \mathbf{r}$ for $n = 1, 2, \dots$. Letting $n \rightarrow \infty$, Theorem 4.3 leads to $\mathbf{a} = \mathbf{r}$. \square

Theorem 4.4 may also be expressed by the statement that the homogeneous linear system

$$(I - P)\mathbf{a} = \mathbf{0}$$

has a unique vector solution \mathbf{a} such that $q_1 + q_2 + \dots + q_k = 1$. Let us apply this technique to the computation of the steady-state vectors for our three examples.

EXAMPLE 4.1 (REVISITED) In Example 4.1, the transition matrix was

$$P = \begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix},$$

so that the linear system $(I - P)\mathbf{a} = \mathbf{0}$ is

$$\begin{bmatrix} .2 & -.3 \\ -.2 & .3 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (4.2)$$

This leads to the single independent equation

$$.2q_1 - .3q_2 = 0$$

or

$$q_1 = 1.5q_2.$$

Thus, setting $q_2 = s$, any solution of (4.2) is of the form $\mathbf{a} = s \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}$

where s is an arbitrary constant. To make \mathbf{a} a probability vector we set $s = 1/(1.5 + 1) = .4$. Consequently,

$$\mathbf{a} = \begin{bmatrix} .6 \\ .4 \end{bmatrix}$$

is the steady-state vector of this regular Markov process. This means that over the long run, 60% of the alumni will give a donation in any one year, and 40% will not. Observe that this agrees with the result obtained numerically in Example 4.1.

EXAMPLE 4.2 (REVISITED) In Example 4.2, the transition matrix was

$$P = \begin{bmatrix} .8 & .3 & .2 \\ .1 & .2 & .6 \\ .1 & .5 & .2 \end{bmatrix},$$

so that the linear system $(I - P)\mathbf{a} = \mathbf{0}$ is

$$\begin{bmatrix} .2 & -.3 & -.2 \\ -.1 & .8 & -.6 \\ -.1 & -.5 & .8 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

The reduced row-echelon form of the coefficient matrix is (verify)

$$\begin{bmatrix} 1 & 0 & -34/13 \\ 0 & 1 & -14/13 \\ 0 & 0 & 0 \end{bmatrix},$$

so that the original linear system is equivalent to the system

$$\begin{aligned} q_1 &= (34/13)q_3 \\ q_2 &= (14/13)q_3. \end{aligned}$$

Setting $q_3 = s$, any solution of the linear system is of the form

$$\mathbf{q} = s \begin{bmatrix} 34/13 \\ 14/13 \\ 1 \end{bmatrix}.$$

To make this a probability vector, we set $s = \frac{1}{\frac{34}{13} + \frac{14}{13} + 1} = \frac{13}{61}$. Thus

the steady-state vector of the system is

$$\mathbf{q} = \begin{bmatrix} 34/61 \\ 14/61 \\ 13/61 \end{bmatrix} = \begin{bmatrix} .5573\dots \\ .2295\dots \\ .2131\dots \end{bmatrix}.$$

This agrees with the result obtained numerically in Example 4.2. The entries of \mathbf{q} give the long-run probabilities that any one car will be returned to location 1, 2, or 3 respectively. If the car rental agency has a fleet of 1000 cars, it should thus design its facilities so that there are at least 558 spaces at location 1, at least 230 spaces at location 2, and at least 214 spaces at location 3.

EXAMPLE 4.3 (REVISITED) We shall not give the details of the calculations, but shall simply state that the unique probability vector solution of the linear system $(I - P)\mathbf{q} = \mathbf{0}$ is

$$\mathbf{q} = \begin{bmatrix} 3/28 \\ 3/28 \\ 3/28 \\ 5/28 \\ 4/28 \\ 3/28 \\ 4/28 \\ 3/28 \end{bmatrix} = \begin{bmatrix} .1071\dots \\ .1071\dots \\ .1071\dots \\ .1785\dots \\ .1428\dots \\ .1071\dots \\ .1428\dots \\ .1071\dots \end{bmatrix}.$$

The entries in this vector indicate the proportion of time the officer spends at each intersection over the long term. Thus, if the objective is for him to spend the same proportion of time at each intersection, then the strategy of random movement with equal probabilities from one intersection to another is not a good one. (See Exercise 4.5.)

EXERCISES

4.1 For the transition matrix $P = \begin{bmatrix} .4 & .5 \\ .6 & .5 \end{bmatrix}$,

(a) calculate $\mathbf{x}^{(n)}$ for $n = 1, 2, 3, 4, 5$ if $\mathbf{x}^{(0)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

(b) state why P is regular and find its steady-state vector.

4.2 For the transition matrix

$$P = \begin{bmatrix} .2 & .1 & .7 \\ .6 & .4 & .2 \\ .2 & .5 & .1 \end{bmatrix},$$

(a) calculate $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, and $\mathbf{x}^{(3)}$ to three decimal places if

$$\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

(b) state why P is regular and find its steady-state vector.

4.3 Find the steady-state vectors of the following regular transition matrices:

$$(a) \begin{bmatrix} 1/3 & 3/4 \\ 2/3 & 1/4 \end{bmatrix} \quad (b) \begin{bmatrix} .81 & .26 \\ .19 & .74 \end{bmatrix} \quad (c) \begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/4 \\ 1/3 & 1/2 & 3/4 \end{bmatrix}$$

4.4 Let P be the transition matrix

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

(a) Show that P is not regular.

(b) Show that as n increases, $P^n \mathbf{x}^{(0)}$ approaches $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ for any initial state vector $\mathbf{x}^{(0)}$.

LEONTIEF CLOSED (INPUT-OUTPUT) MODEL

We begin with a simple example, and then proceed to the general theory of the model.

EXAMPLE 5.1 Three homeowners — a carpenter, an electrician, and a plumber — mutually agree to make repairs in their three homes. They agree to work a total of ten days each according to the following schedule:

	Work performed by		
	carpenter	electrician	plumber
Days of work in home of carpenter	2	1	6
Days of work in home of electrician	4	5	1
Days of work in home of plumber	4	4	3

For tax purposes, they must report and pay each other a reasonable daily wage, even for the work each does on his own home. Their normal daily wages are in the \$60 - \$80 range, but they agree to adjust their respective daily wages so that each homeowner will come out even; that is, so that the total amount paid out by each is the same as the total amount each receives. Let us set

$$\begin{aligned} p_1 &= \text{daily wage of carpenter} \\ p_2 &= \text{daily wage of electrician} \\ p_3 &= \text{daily wage of plumber.} \end{aligned}$$

To satisfy the "equilibrium" condition that each homeowner comes out even, we require that

$$\text{total expenditures} = \text{total income}$$

for each of the homeowners for the ten day period. For example, the carpenter pays a total of $2p_1 + p_2 + 6p_3$ for the repairs in his own home, and receives a total income of $10p_1$ for the repairs that he performs on all three homes. Equating these two expressions then gives the first of the following three equations:

$$\begin{aligned} 2p_1 + p_2 + 6p_3 &= 10p_1 \\ 4p_1 + 5p_2 + p_3 &= 10p_2 \\ 4p_1 + 4p_2 + 3p_3 &= 10p_3. \end{aligned}$$

The remaining two equations are the equilibrium equations for the electrician and the plumber. Dividing these equations by ten and rewriting them in matrix form yields

$$\begin{bmatrix} .2 & .1 & .6 \\ .4 & .5 & .1 \\ .4 & .4 & .3 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}. \quad (5.1)$$

Equation (5.1) can be rewritten as a homogeneous system by subtracting the lefthand side from the righthand side to obtain

$$\begin{bmatrix} .8 & -.1 & -.6 \\ -.4 & .5 & -.1 \\ -.4 & -.4 & .7 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

The solution of this homogeneous system is easily found to be (verify)

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = s \begin{bmatrix} 31 \\ 32 \\ 36 \end{bmatrix},$$

where s is an arbitrary constant. This constant is a scale factor which the homeowners may choose for their convenience. For example, they may set $s = 2$ so that the corresponding daily wages — \$62, \$64 and \$72 — fall in the \$60 - \$80 range.

This example illustrates the salient features of the Leontief input-output model of a closed economy. In the basic equation (5.1), each column sum of the coefficient matrix is one, corresponding to the fact that each of the homeowners' "output" of labor is completely distributed among these same homeowners in the proportions given by the entries in the column. Our problem was to determine suitable "prices" for these outputs to put the system in equilibrium; that is so that each homeowner's total expenditures equal his total income.

In the general model, we have an economic system consisting of a finite number of "industries", which we number as industries 1, 2, ..., k . Over some fixed period of time, each industry produces an "output" of some good or service which is completely utilized in a predetermined manner by the k industries. An important problem is to find suitable "prices" to be charged for these k outputs so that for each industry total expenditures equal total income. Such a price structure represents an equilibrium position for the economy.

For the fixed time period in question, let us set

p_i = price charged by the i th industry for its total output,

e_{ij} = fraction of the total output of the j th industry purchased by the i th industry.

for $i, j = 1, 2, \dots, k$. By definition, we have

- (i) $p_i \geq 0$, $i = 1, 2, \dots, k$.
- (ii) $e_{ij} \geq 0$, $i, j = 1, 2, \dots, k$.
- (iii) $e_{1j} + e_{2j} + \dots + e_{kj} = 1$, $j = 1, 2, \dots, k$.

With these quantities, we form the *price vector*

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{bmatrix}$$

and the *exchange or input-output matrix*

$$E = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1k} \\ e_{21} & e_{22} & \cdots & e_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ e_{k1} & e_{k2} & \cdots & e_{kk} \end{bmatrix}$$

Condition (iii) above expresses the fact that all of the column sums of the exchange matrix are one.

As in the example, in order that the expenditures of each industry be equal to its income, the following matrix equation must be satisfied (see (5.1)):

$$E\mathbf{p} = \mathbf{p} \quad (5.2)$$

or

$$(I - E)\mathbf{p} = \mathbf{0}. \quad (5.3)$$

Equation (5.3) is a homogeneous linear system for the price vector \mathbf{p} . It will have a nontrivial solution if and only if the determinant of its coefficient matrix $I - E$ is zero. In Exercise 5.6 we ask the reader to show that this is the case for any exchange matrix E . Thus, (5.3) always has nontrivial solutions for the price vector \mathbf{p} .

Actually, for our economic model to make sense we need more than just the fact that (5.3) has nontrivial solutions for \mathbf{p} . We also need that the prices p_i of the k outputs are nonnegative numbers. We express this condition as $\mathbf{p} \geq \mathbf{0}$. (In general, if A is any vector or matrix, the notation $A \geq \mathbf{0}$ means that every entry of A is nonnegative, and the notation $A > \mathbf{0}$ means that every entry of A is positive. Similarly, $A > B$ denotes $A - B > \mathbf{0}$, and $A \geq B$ denotes $A - B \geq \mathbf{0}$.) To show that (5.3) has a nontrivial solution for which $\mathbf{p} > \mathbf{0}$ is a bit more difficult than showing merely that some nontrivial solution exists. But it is true, and we state this fact without proof in the following theorem:

THEOREM 5.1 *If E is an exchange matrix, then $E\mathbf{p} = \mathbf{p}$ always has a nontrivial solution \mathbf{p} whose entries are nonnegative.*

Let us consider a few simple examples of this theorem.

EXAMPLE 5.2

Let

$$E = \begin{bmatrix} 1/2 & 0 \\ 1/2 & 1 \end{bmatrix}$$

Then $(I - E)\mathbf{p} = \mathbf{0}$ is

$$\begin{bmatrix} 1/2 & 0 \\ -1/2 & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which has the general solution

$$\mathbf{p} = s \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

where s is an arbitrary constant. We then have nontrivial solutions $\mathbf{p} \geq \mathbf{0}$ for any $s > 0$.

EXAMPLE 5.3

Let

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

Then $(I - E)p = 0$ has the general solution

$$p = s \begin{bmatrix} 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

where s and t are independent arbitrary constants. Nontrivial solutions $p \geq 0$ then result from any $s \geq 0$ and $t \geq 0$, not both zero.

Example 5.2 indicates that in some situations one of the prices must be zero in order to satisfy the equilibrium condition. Example 5.3 indicates that there may be several linearly independent price structures available. Neither of these situations describes a truly interdependent economic structure. The following theorem gives sufficient conditions for both cases to be excluded.

THEOREM 5.2 *Let E be an exchange matrix such that for some positive integer m , all of the entries of E^m are positive. Then there is exactly one linearly independent solution of $(I - E)p = 0$, and it may be chosen so that all of its entries are positive.*

We shall not give a proof of this theorem. The reader who has read Chapter 4 on Markov Chains may observe that this theorem is essentially the same as Theorem 4.3. What we are calling exchange matrices in this chapter were called transition or Markov matrices in Chapter 4.

EXAMPLE 5.4 The exchange matrix in Example 5.1 was

$$E = \begin{bmatrix} .2 & .1 & .6 \\ .4 & .5 & .1 \\ .4 & .4 & .3 \end{bmatrix}.$$

Since $E > 0$, the condition $E^m > 0$ in Theorem 5.2 is satisfied for $m = 1$. Consequently, we are guaranteed that there is exactly one linearly independent solution to $(I - E)p = 0$, and it may be chosen so that $p > 0$. In that example, we found that

$$p = \begin{bmatrix} 31 \\ 32 \\ 36 \end{bmatrix}$$

is such a solution.

LEONTIEF OPEN (PRODUCTION) MODEL

In contrast with the closed model in which the outputs of k industries are distributed only among themselves, the open model attempts to satisfy an outside demand for the outputs. Portions of these outputs may still be distributed among the industries themselves, to keep them operating, but there is to be some excess, some net production, with which to satisfy the outside demand. In the closed model, the outputs of the industries were fixed, and our objective was to determine prices for these outputs so that the equilibrium condition, that expenditures equal incomes, was satisfied. In the open model, it is the prices which are fixed, and our objective will be to determine levels of the outputs of the industries needed to satisfy the outside demand. We shall measure the levels of the outputs in terms of their economic values using the fixed prices. To be precise, over some fixed period of time let

x_i = monetary value of the total output of the i th industry,

d_i = monetary value of the output of the i th industry needed to satisfy the outside demand,

c_{ij} = monetary value of the output of the i th industry needed by the j th industry to produce one unit of monetary value of its own output.

With these quantities, we define the *production vector*

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix},$$

the *demand vector*

$$d = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{bmatrix},$$

and the *consumption matrix*

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1k} \\ c_{21} & c_{22} & \cdots & c_{2k} \\ \vdots & \vdots & & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kk} \end{bmatrix}$$

By their nature, we have that

$$x \geq 0, \quad d \geq 0, \quad \text{and} \quad C \geq 0.$$

From the definition of c_{ij} and x_j , it can be seen that the quantity

$$c_{i1}x_1 + c_{i2}x_2 + \cdots + c_{ik}x_k$$

is the value of the output of the i th industry needed by all k industries to produce a total output specified by the production vector x . Since this quantity is simply the i th entry of the column vector Cx , we can further say that the i th entry of the column vector

$$x - Cx$$

is the value of the excess output of the i th industry available to satisfy the outside demand. The value of the outside demand for the output of the i th industry is the i th entry of the demand vector d . Consequently, we are led to the following equation

$$x - Cx = d$$

or

$$(I - C)x = d \quad (5.4)$$

for the demand to be exactly met, without any surpluses or shortages. Thus, given C and d , our objective is to find a production vector $x \geq 0$ which satisfies Equation (5.4).

EXAMPLE 5.5 A town has three main industries: a coal-mining operation, an electric power generating plant, and a local railroad. To mine \$1 of coal, the mining operation must purchase \$.25 of electricity to run its equipment and \$.25 of transportation for its shipping needs. To produce \$1 of electricity, the generating plant requires \$.65 of coal for fuel, \$.05 of its own electricity to run auxiliary equipment, and \$.05 of transportation. To provide \$1 of transportation, the railroad requires \$.55 of coal for fuel, and \$.10 of electricity for its auxiliary equipment. In a certain week, the

coal-mining operation receives orders for \$50,000 of coal from outside the town, and the generating plant receives orders for \$25,000 of electricity from outside. There is no outside demand for the local railroad. How much must each of the three industries produce in that week to exactly satisfy their own demand and the outside demand?

SOLUTION

For the one week period, let

- x_1 = value of total output of coal-mining operation,
- x_2 = value of total output of power generating plant,
- x_3 = value of total output of local railroad.

From the information supplied, the consumption matrix of the system is

$$C = \begin{bmatrix} 0 & .65 & .55 \\ .25 & .05 & .10 \\ .25 & .05 & 0 \end{bmatrix}$$

The linear system $(I - C)x = d$ is then

$$\begin{bmatrix} 1.00 & -.65 & -.55 \\ -.25 & .95 & -.10 \\ -.25 & -.05 & 1.00 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 50,000 \\ 25,000 \\ 0 \end{bmatrix}$$

The coefficient matrix on the left is invertible, and the solution is given by

$$x = (I - C)^{-1}d = \frac{1}{503} \begin{bmatrix} 756 & 542 & 470 \\ 220 & 690 & 190 \\ 200 & 170 & 630 \end{bmatrix} \begin{bmatrix} 50,000 \\ 25,000 \\ 0 \end{bmatrix} = \begin{bmatrix} 102,087 \\ 56,163 \\ 28,330 \end{bmatrix}$$

Thus, the total output of the coal-mining operation should be \$102,087, the total output of the power generating plant should be \$56,163, and the total output of the railroad should be \$28,330.

Let us reconsider Eq. (5.4):

$$(I - C)x = d.$$

If the square matrix $I - C$ is invertible, we may write

$$x = (I - C)^{-1}d. \quad (5.5)$$

In addition, if the matrix $(I - C)^{-1}$ has only nonnegative entries, then we are guaranteed that for any $d > 0$, Eq. (5.5) determines a unique nonnegative solution for x . This is a particularly desirable situation since it means that any outside demand can be met. The terminology used to describe this case is given in the following definition:

DEFINITION 5.1 A consumption matrix C is said to be productive if $(I - C)^{-1}$ exists and $(I - C)^{-1} \geq 0$.

We shall now consider some simple criteria which will guarantee that a consumption matrix is productive. The first is given in the following theorem:

THEOREM 5.3 A consumption matrix C is productive if and only if there is some production vector $x \geq 0$ such that $x > Cx$.

(The proof is outlined in Exercise 5.8.) The condition $x > Cx$ means that there is some production schedule possible such that each industry produces more than it consumes.

Theorem 5.3 has two interesting corollaries. Suppose all of the row sums of C are less than one. Then if

$$x = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

Cx is a column vector whose entries are these row sums. Thus $x > Cx$, and the condition of Theorem 5.3 is satisfied. We state this as

COROLLARY 5.1 A consumption matrix is productive if each of its row sums is less than one.

As we ask the reader to show in Exercise 5.7, this corollary leads to

COROLLARY 5.2 A consumption matrix is productive if each of its column sums is less than one.

Recalling the definition of the entries of the consumption matrix C , we see that the j th column sum of C is the total value of the outputs of all k industries needed to produce one unit of value of output of the j th industry. The j th industry is thus said to be profitable if that j th column sum is less than one. In other words, Corollary 5.2 says that a consumption matrix is productive if all k industries in the economic system are profitable.

EXAMPLE 5.6 The consumption matrix in Example 5.5 was

$$C = \begin{bmatrix} 0 & .65 & .55 \\ .25 & .05 & .10 \\ .25 & .05 & 0 \end{bmatrix}.$$

All three column sums in this matrix are less than one, and so all three industries are profitable. Consequently, by Corollary 5.2, the consumption matrix C is productive. This can also be seen in the calculations in Example 5.5 since $(I - C)^{-1}$ is nonnegative.

EXERCISES

5.1 For the following exchange matrices, find nonnegative price vectors which satisfy the equilibrium condition (5.3).

$$(a) \begin{bmatrix} 1/2 & 1/3 \\ 1/2 & 2/3 \end{bmatrix} \quad (b) \begin{bmatrix} 1/2 & 0 & 1/2 \\ 1/3 & 0 & 1/2 \\ 1/6 & 1 & 0 \end{bmatrix} \quad (c) \begin{bmatrix} .35 & .50 & .30 \\ .25 & .20 & .30 \\ .40 & .30 & .40 \end{bmatrix}$$

5.2 Using Theorem 5.3 and its corollaries, show that each of the following consumption matrices is productive.

$$(a) \begin{bmatrix} .8 & .1 \\ .3 & .6 \end{bmatrix} \quad (b) \begin{bmatrix} .70 & .30 & .25 \\ .20 & .40 & .25 \\ .05 & .15 & .25 \end{bmatrix} \quad (c) \begin{bmatrix} .7 & .3 & .2 \\ .1 & .4 & .3 \\ .2 & .4 & .1 \end{bmatrix}$$

5.3 Using Theorem 5.2, show that there is only one linearly independent price vector for the closed economic system with exchange matrix

$$E = \begin{bmatrix} 0 & .2 & .5 \\ 1 & .2 & .5 \\ 0 & .6 & 0 \end{bmatrix}.$$

5.4 Three neighbors have backyard vegetable gardens. Neighbor *A* grows tomatoes, neighbor *B* grows corn, and neighbor *C* grows lettuce. They agree to divide their crops among themselves as follows: *A* gets 1/2 of the tomatoes, 1/3 of the corn, and 1/4 of the lettuce. *B* gets 1/3 of the tomatoes, 1/3 of the corn, and 1/4 of the lettuce. *C* gets 1/6 of the tomatoes, 1/3 of the corn, and 1/2 of the lettuce. What prices should the neighbors assign to their respective crops if the equilibrium condition of a closed economy is to be satisfied, and if the lowest-priced crop is to have a price of \$100?

5.5 Three engineers — a civil engineer (CE), an electrical engineer (EE), and a mechanical engineer (ME) — each have a consulting firm. The consulting they do is of a multidisciplinary nature, and so they buy a portion of each others' services. For each \$1 of consulting the CE does, he buys \$.10 of the EE's services and \$.30 of the ME's services. For each \$1 of consulting the EE does, he buys \$.20 of the CE's services and \$.40 of the ME's services. And for each \$1 of consulting the ME does, he buys \$.30 of the CE's services and \$.40 of the EE's services. In a certain week, the CE receives outside consulting orders of \$500, the EE receives outside consulting orders of \$700, and the ME receives outside consulting orders of \$600. What dollar amount of consulting does each engineer perform in that week?

5.6 Using the fact that the column sums of an exchange matrix *E* are all one, show that the column sums of *I - E* are zero. From this, show that *I - E* has zero determinant, and so $(I - E)p = 0$ has non-trivial solutions for *p*.

5.7 Show that Corollary 5.2 follows from Corollary 5.1. Hint: use the fact that $(A^t)^{-1} = (A^{-1})^t$ for any invertible matrix *A*.

5.8 Prove Theorem 5.3 as follows:

I. Prove the "only if" part of theorem; i.e. show that if *C* is a productive consumption matrix, then there is a vector $x \geq 0$ such that $x > Cx$.

II. Prove the "if" part of theorem as follows:

Step 1. Show that if there is a vector $x^* \geq 0$ such that $Cx^* < x^*$, then $x^* > 0$.

Step 2. Show that there is a number λ such that $0 < \lambda < 1$ and $Cx^* < \lambda x^*$.

Step 3. Show that $C^n x^* < \lambda^n x^*$ for $n = 1, 2, \dots$.

Step 4. Show that $C^n \rightarrow 0$ as $n \rightarrow \infty$.

Step 5. By direct verification, show that

$$(I - C)(I + C + C^2 + \dots + C^{n-1}) = I - C^n$$

for $n = 1, 2, \dots$.

Step 6. By letting $n \rightarrow \infty$ in step 5, show that the matrix infinite sum

$$S = I + C + C^2 + \dots$$

exists and that $(I - C)S = I$.

Step 7. Show that $S \geq 0$ and that $S = (I - C)^{-1}$.

Step 8. Conclude from Definition 5.1 that *C* is a productive consumption matrix.

Forest Management

A matrix model for the management of a forest whose trees are grouped into height classes is presented. The optimal sustainable yield of a periodic harvest is calculated when the trees of different height classes can have different economic values.

PREREQUISITES: Matrix multiplication and addition

INTRODUCTION

In this chapter, we shall introduce a simplified model for the sustainable harvesting of a forest whose trees are classified by height. The height of a tree is assumed to determine its economic value when it is cut down and sold. Initially, there is a distribution of trees of various heights. The forest is then allowed to grow for a certain period of time, after which some of the trees of various heights are harvested. But the trees left unharvested are to be of the same height configuration as the original forest, so that the harvest is sustainable. As we shall see, there are many such sustainable harvesting procedures. We want to find one for which the total economic value of all the trees removed is as large as possible. This determines the *optimal sustainable yield* of the forest, and is the largest yield which can be attained continually without depleting the forest.

THE MODEL

Suppose a harvester has a forest of Douglas fir trees which he wants to sell as Christmas trees year after year. Every December he cuts down some of the trees to be sold. For each tree he cuts down, he plants a seedling in its place. In this way, the total number of trees in the forest is always the same. (In this simplified model, we shall not take into account trees which die between harvests. We assume that every seedling planted survives and grows until it is harvested.)

In the market place, trees of different heights have different economic values. Suppose there are n different price classes corresponding to certain height intervals as follows:

Class	Value in Dollars	Height Interval
1 (seedlings)	none	$[0, h_1)$
2	p_2	$[h_1, h_2)$
3	p_3	$[h_2, h_3)$
\vdots	\vdots	\vdots
$n-1$	p_{n-1}	$[h_{n-2}, h_{n-1})$
n	p_n	$[h_{n-1}, \infty)$

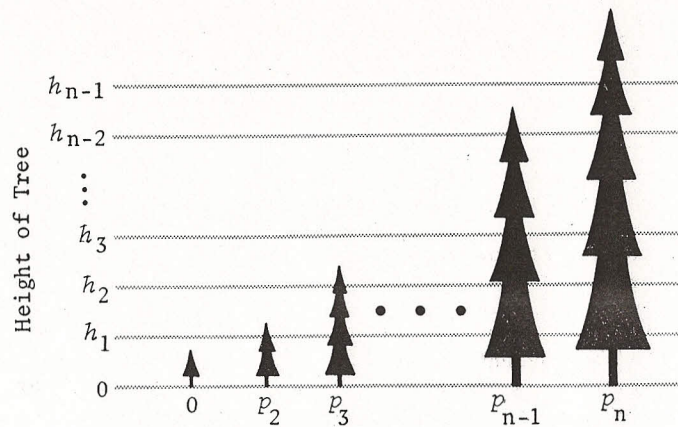


Figure 6.1

The first class consists of seedlings with heights in the interval $[0, h_1)$, and these seedlings are of no economic value. The n -th class consists of trees with heights greater than or equal to h_{n-1} .

Let $x_i (i = 1, 2, \dots, n)$ be the number of trees within the i -th class which remain after each harvest. We form a column vector with these numbers

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

and call it the *nonharvest vector*. For a sustainable harvesting policy, the forest is to be returned after each harvest to the fixed configuration given by the nonharvest vector x . Part of our problem will be to find those nonharvest vectors x for which sustainable harvesting is possible.

Since the total number of trees in the forest is fixed, let us set

$$x_1 + x_2 + \dots + x_n = s \tag{6.1}$$

where s is predetermined by the amount of land available and the amount of space each tree requires. Referring to Fig. 6.2, we have the following situation: The forest configuration is given by the vector x after each harvest. Between harvests, the trees grow and produce a new forest configuration before each harvest. A certain number of trees are removed from each class at the harvest. Finally, a seedling is planted in place of each tree removed, to return the forest again to the configuration x .

Let us first consider the growth of the forest between harvests. During this period, a tree in the i -th class may grow and move up to a higher height class. Or its growth may be retarded for some reason, and it will remain in the same class. We consequently define the following growth parameters, g_i , for $i = 1, 2, \dots, n-1$:

g_i = the fraction of trees in the i -th class that grow into the $(i+1)$ -st class during a growth period.

For simplicity, we shall assume that a tree can move at most one height class upward in one growth period. With this assumption we have

$1 - g_i$ = the fraction of trees in the i -th class that remain in the i -th class during a growth period.

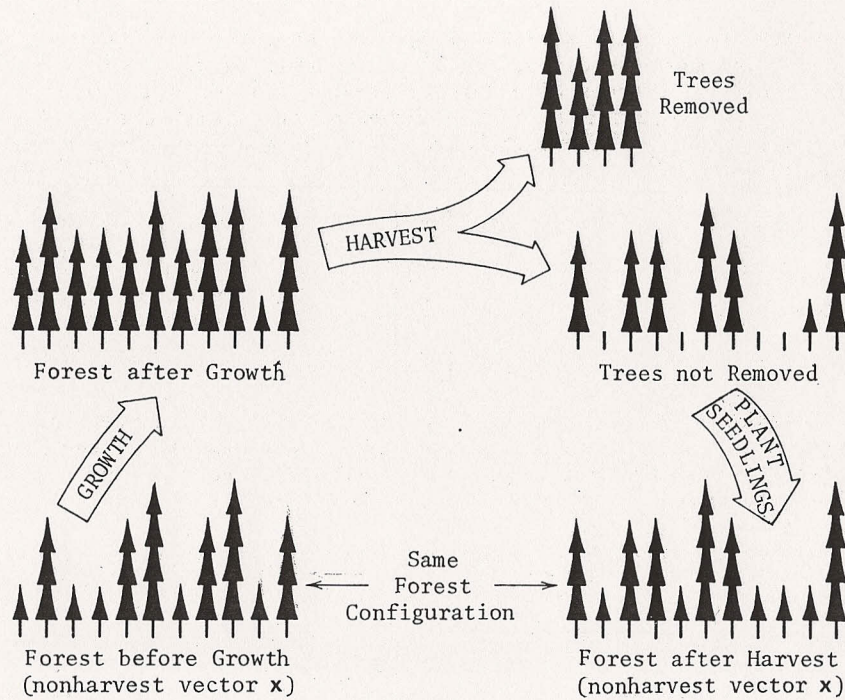


Figure 6.2

With these $n-1$ growth parameters, we form the following $n \times n$ growth matrix:

$$G = \begin{bmatrix} 1-g_1 & 0 & 0 & \dots & 0 \\ g_1 & 1-g_2 & 0 & \dots & 0 \\ 0 & g_2 & 1-g_3 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1-g_{n-1} \\ 0 & 0 & 0 & \dots & g_{n-1} & 1 \end{bmatrix} \quad (6.2)$$

Since the entries of the vector x are the numbers of trees in the n classes before the growth period, the reader can verify that the entries of the vector

$$Gx = \begin{bmatrix} (1-g_1)x_1 \\ g_1x_1 + (1-g_2)x_2 \\ g_2x_2 + (1-g_3)x_3 \\ \vdots \\ g_{n-2}x_{n-2} + (1-g_{n-1})x_{n-1} \\ g_{n-1}x_{n-1} + x_n \end{bmatrix} \quad (6.3)$$

are the numbers of trees in the n classes after the growth period. Suppose during the harvest we remove y_i ($i=1, 2, \dots, n$) trees from the i -th class. We shall call the column vector

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

the harvest vector. Thus, a total of

$$y_1 + y_2 + \dots + y_n$$

trees are removed at each harvest. This is also the total number of trees added to the first class (the new seedlings) after each harvest. If we define the following $n \times n$ replacement matrix

$$R = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} \quad (6.4)$$

then the column vector

$$Ry = \begin{bmatrix} y_1 + y_2 + \dots + y_n \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (6.5)$$

specifies the configuration of trees planted after each harvest.

At this point, we are ready to write the following equation which characterizes a sustainable harvesting policy:

$$\begin{pmatrix} \text{configuration} \\ \text{at end of} \\ \text{growth period} \end{pmatrix} - \begin{pmatrix} \text{harvest} \end{pmatrix} + \begin{pmatrix} \text{new seedling} \\ \text{replacement} \end{pmatrix} = \begin{pmatrix} \text{configuration} \\ \text{at beginning of} \\ \text{growth period} \end{pmatrix}$$

or mathematically

$$Gx - y + Ry = x.$$

This equation can be rewritten as

$$(I - R)y = (G - I)x, \quad (6.6)$$

or, more fully,

$$\begin{bmatrix} 0 & -1 & -1 & \cdots & -1 & -1 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} -g_1 & 0 & 0 & \cdots & 0 & 0 \\ g_1 & -g_2 & 0 & \cdots & 0 & 0 \\ 0 & g_2 & -g_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -g_{n-1} & 0 \\ 0 & 0 & 0 & \cdots & g_{n-1} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}.$$

We shall refer to Eq. (6.6) as the *sustainable harvesting condition*. Any vectors x and y with nonnegative entries, and such that $x_1 + x_2 + \cdots + x_n = s$, which satisfy this matrix equation determine a sustainable harvesting policy for the forest. Let us note that if $y_1 > 0$, then the harvester is removing seedlings, of no economic value, and then replacing them with new seedlings. Since there is no point in doing this, we shall assume

$$y_1 = 0. \quad (6.7)$$

With this assumption, it may be verified that (6.6) is the matrix form of the following set of equations:

$$\begin{aligned} y_2 + y_3 + \cdots + y_n &= g_1 x_1 \\ y_2 &= g_1 x_1 - g_2 x_2 \\ y_3 &= g_2 x_2 - g_3 x_3 \\ &\vdots \\ y_{n-1} &= g_{n-2} x_{n-2} - g_{n-1} x_{n-1} \\ y_n &= g_{n-1} x_{n-1}. \end{aligned} \quad (6.8)$$

Notice that the first of Eqs. (6.8) is the sum of the remaining $n-1$ equations.

Since we must have $y_i \geq 0$ for $i = 2, 3, \dots, n$, Eqs. (6.8) require that

$$g_1 x_1 \geq g_2 x_2 \geq \cdots \geq g_{n-1} x_{n-1} \geq 0. \quad (6.9)$$

Conversely, if x is a column vector with nonnegative entries which satisfy Eq. (6.9), then Eqs. (6.7) and (6.8) define a column vector y with nonnegative entries. Furthermore, x and y then satisfy the sustainable harvesting condition (6.6). In other words, a necessary and sufficient condition that a nonnegative column vector x determine a forest configuration which is capable of sustainable harvesting is that its entries satisfy (6.9).

OPTIMAL SUSTAINABLE YIELD

Since we remove y_i trees from the i -th class ($i = 2, 3, \dots, n$) and since each tree in the i -th class has an economic value of p_i , the total yield of the harvest, Yld , is given by

$$Yld = p_2 y_2 + p_3 y_3 + \cdots + p_n y_n. \quad (6.10)$$

Using (6.8), we may substitute for the y_i 's in (6.10) to obtain

$$Yld = p_2 g_1 x_1 + (p_3 - p_2) g_2 x_2 + \cdots + (p_n - p_{n-1}) g_{n-1} x_{n-1}. \quad (6.11)$$

Combining Eqs. (6.11), (6.1), and (6.9), we may now state the problem of maximizing the yield of the forest over all possible sustainable harvesting policies as

Problem Find nonnegative x_1, x_2, \dots, x_n which maximize

$$Yld = p_2 g_1 x_1 + (p_3 - p_2) g_2 x_2 + \dots + (p_n - p_{n-1}) g_{n-1} x_{n-1}$$

subject to

$$x_1 + x_2 + \dots + x_n = s$$

and

$$g_1 x_1 \geq g_2 x_2 \geq \dots \geq g_{n-1} x_{n-1} \geq 0.$$

As formulated above, this problem belongs to the field of Linear Programming. (See Chapter 13 for an introduction to this field.) However, we shall only need the following result from Linear Programming theory applied to this problem:

The optimal sustainable yield is achieved by harvesting all of the trees from one particular height class and none of the trees from any other height class.

Below, we prove this result without Linear Programming theory by actually exhibiting such a sustainable harvesting policy.

Let us first set

Yld_k = yield obtained by harvesting all of the k -th class and none of the other classes.

The largest value of Yld_k for $k = 2, 3, \dots, n$ will then be the optimal sustainable yield, and the corresponding value of k will be the class which should be completely harvested to attain the optimal sustainable yield. Since no class but the k -th is harvested, we have

$$y_2 = y_3 = \dots = y_{k-1} = y_{k+1} = \dots = y_n = 0. \quad (6.12)$$

In addition, since all of the k -th class is harvested, no trees are left unharvested in the k -th class, and no trees are ever present in the height classes above the k -th class. Thus,

$$x_k = x_{k+1} = \dots = x_n = 0. \quad (6.13)$$

Substitution of (6.12) and (6.13) into the sustainable harvesting condition (6.8) then gives

$$\begin{aligned} y_k &= g_1 x_1 \\ 0 &= g_1 x_1 - g_2 x_2 \\ 0 &= g_2 x_2 - g_3 x_3 \\ &\vdots \\ 0 &= g_{k-2} x_{k-2} - g_{k-1} x_{k-1} \\ y_k &= g_{k-1} x_{k-1}. \end{aligned} \quad (6.14)$$

Equations (6.14) may also be written as

$$y_k = g_1 x_1 = g_2 x_2 = \dots = g_{k-1} x_{k-1}, \quad (6.15)$$

from which it follows that

$$\begin{aligned} x_2 &= g_1 x_1 / g_2 \\ x_3 &= g_1 x_1 / g_3 \\ &\vdots \\ x_{k-1} &= g_1 x_1 / g_{k-1}. \end{aligned} \quad (6.16)$$

If we substitute Eqs. (6.13) and (6.16) into (6.1),

$$x_1 + x_2 + \dots + x_n = s,$$

we may solve for x_1 and obtain

$$x_1 = \frac{s}{1 + \frac{g_1}{g_2} + \frac{g_1}{g_3} + \dots + \frac{g_1}{g_{k-1}}}. \quad (6.17)$$

For the yield Yld_k , we combine Eqs. (6.10), (6.12), (6.15) and (6.17) to obtain

$$\begin{aligned}
 Yld_k &= p_2 y_2 + p_3 y_3 + \dots + p_n y_n \\
 &= p_k y_k \\
 &= p_k g_1^x \\
 &= \frac{p_k s}{\frac{1}{g_1} + \frac{1}{g_2} + \dots + \frac{1}{g_{k-1}}} \quad (6.18)
 \end{aligned}$$

Equation (6.18) determines Yld_k in terms of the known growth and economic parameters for any $k=2, 3, \dots, n$. Thus, the optimal sustainable yield is found as follows:

THEOREM 6.1 *The optimal sustainable yield is the largest value of*

$$\frac{p_k s}{\frac{1}{g_1} + \frac{1}{g_2} + \dots + \frac{1}{g_{k-1}}}$$

for $k=2, 3, \dots, n$. The corresponding value of k is the number of the class which is completely harvested.

In Exercise (6.4), we ask the reader to show that the nonharvest vector x for the optimal sustainable yield is

$$x = \frac{s}{\frac{1}{g_1} + \frac{1}{g_2} + \dots + \frac{1}{g_{k-1}}} \begin{bmatrix} 1/g_1 \\ 1/g_2 \\ \vdots \\ 1/g_{k-1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (6.19)$$

Theorem 6.1 implies that it is not necessarily the highest-priced class of trees which should be totally cropped. The growth parameters, g_i , must also be taken into account to determine the optimal sustainable yield.

EXAMPLE 6.1 For a Scots Pine forest in Scotland with a growth period of six years, the following growth matrix was found (M. B. Usher, "A Matrix Approach to the Management of Renewable Resources, with Special Reference to Selection Forests," *Journal of Applied Ecology*, Vol. 3, 1966, pages 355-367):

$$G = \begin{bmatrix} .72 & 0 & 0 & 0 & 0 & 0 \\ .28 & .69 & 0 & 0 & 0 & 0 \\ 0 & .31 & .75 & 0 & 0 & 0 \\ 0 & 0 & .25 & .77 & 0 & 0 \\ 0 & 0 & 0 & .23 & .63 & 0 \\ 0 & 0 & 0 & 0 & .37 & 1.00 \end{bmatrix}$$

Suppose the prices of trees in the five tallest height classes are

$$p_2 = \$50, \quad p_3 = \$100, \quad p_4 = \$150, \quad p_5 = \$200, \quad p_6 = \$250.$$

Which class should be completely harvested to obtain the optimal sustainable yield, and what is the yield?

SOLUTION From the matrix G we have that

$$g_1 = .28, \quad g_2 = .31, \quad g_3 = .25, \quad g_4 = .23, \quad g_5 = .37.$$

Equation (6.18) then gives

$$Yld_2 = 50s / (.28^{-1}) = 14.0s$$

$$Yld_3 = 100s / (.28^{-1} + .31^{-1}) = 14.7s$$

$$Yld_4 = 150s / (.28^{-1} + .31^{-1} + .25^{-1}) = 13.9s$$

$$Yld_5 = 200s / (.28^{-1} + .31^{-1} + .25^{-1} + .23^{-1}) = 13.2s$$

$$Yld_6 = 250s / (.28^{-1} + .31^{-1} + .25^{-1} + .23^{-1} + .37^{-1}) = 14.0s.$$

We see that Yld_3 is the largest of these five quantities, so from Theorem 6.1 the third class should be completely harvested every six years to maximize the sustainable yield. The corresponding optimal sustainable yield is $\$14.7s$, where s is the total number of trees in the forest.

EXERCISES

- 6.1 A certain forest is divided into three height classes and has a growth matrix between harvests given by

$$G = \begin{bmatrix} 1/2 & 0 & 0 \\ 1/2 & 1/3 & 0 \\ 0 & 2/3 & 1 \end{bmatrix}.$$

If the price of trees in the second class is \$30 and the price of trees in the third class is \$50, which class should be completely harvested to attain the optimal sustainable yield? What is the optimal yield if there are 1,000 trees in the forest?

- 6.2 In Example 6.1, to what level must the price of trees in the fifth class rise so that class is the one to completely harvest in order to attain the optimal sustainable yield?
- 6.3 In Example 6.1, what must the ratio of the prices $p_2 : p_3 : p_4 : p_5 : p_6$ be in order that the yields Yld_k , $k = 2, 3, 4, 5, 6$, all be the same? (In this case, any sustainable harvesting policy will produce the same optimal sustainable yield.)
- 6.4 Derive Eq. (6.19) for the nonharvest vector x corresponding to the optimal sustainable harvesting policy described in Theorem 6.1.
- 6.5 For the optimal sustainable harvesting policy described in Theorem 6.1, how many trees are removed from the forest during each harvest?
- 6.6 If all of the growth parameters g_1, g_2, \dots, g_{n-1} in the growth matrix G are equal, what should the ratio of the prices $p_2 : p_3 : \dots : p_n$ be in order that any sustainable harvesting policy be an optimal sustainable harvesting policy? (See Exercise 6.3.)

Equilibrium Temperature Distributions

7

The equilibrium temperature distribution within a trapezoidal plate is found when the temperature around the edges of the plate is specified. The problem is reduced to solving a linear system of equations and an iterative technique for solving the system is described. A "random walk" approach to the problem is also discussed.

PREREQUISITES: Linear systems
Matrices
Intuitive understanding of limits

INTRODUCTION

Suppose we are given a thin trapezoidal plate (Fig. 7.1(a)), whose two faces are insulated from heat. Suppose we are also given the temperature along the four edges of the plate. For example, let the temperature be constant on each edge with values of 0° , 0° , 1° , and 2° , as in the diagram. After a period of time, the temperature inside the plate will stabilize. Our objective in this chapter is

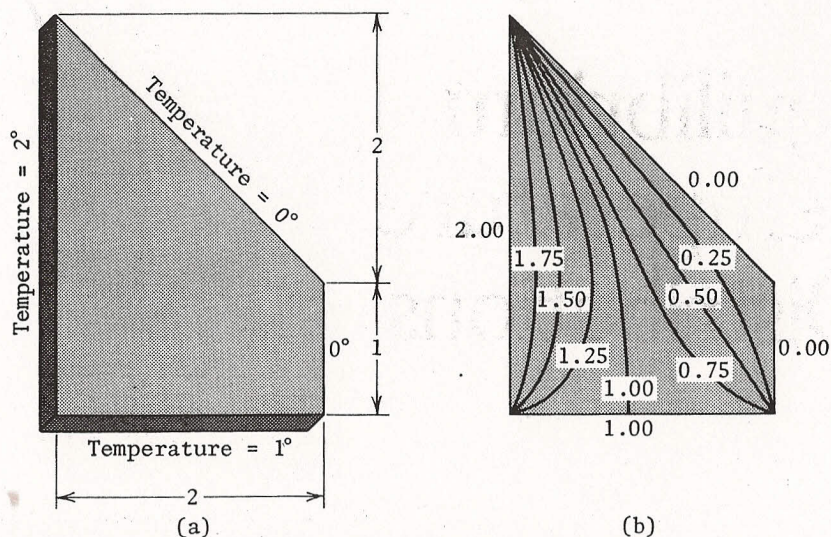


Figure 7.1

to determine this equilibrium temperature distribution at the points inside the plate. As we shall see, the interior equilibrium temperature is completely determined by the *boundary data*, that is, the temperature along the edges of the plate.

The equilibrium temperature distribution can be visualized by the use of curves which connect points of equal temperature. Such curves are called *isotherms* of the temperature distribution. In Fig. 7.1(b) we have sketched a few isotherms using information we derive later in the chapter.

Though all of our calculations will be for the trapezoidal plate illustrated, our techniques generalize easily to a plate of any shape. They also generalize to the problem of finding the temperature within a three-dimensional body. In fact, our "plate" could be the cross-section of some solid object if the flow of heat perpendicular to the cross-section is negligible. For example, Fig. 7.1 could represent the cross-section of a long dam. The dam is exposed to three different temperatures: the temperature of the ground at its base, the temperature of the water on one side, and the temperature of the air on the other side. A knowledge of the temperature distribution inside the dam is necessary to determine the thermal stresses which it is subjected to.

In the next section, we begin with a certain thermodynamic principle which characterizes the temperature distribution we are seeking.

THE MEAN-VALUE PROPERTY

There are many different ways to obtain a mathematical model for our problem. The approach we shall use is based on the following property of equilibrium temperature distributions:

The Mean-Value Property

Let a plate be in thermal equilibrium and let P be a point inside the plate. Then if C is any circle with center at P which is completely contained in the plate, the temperature at P is the average value of the temperature on the circle.

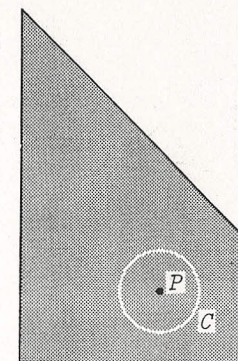


Figure 7.2

This property is a consequence of certain basic laws of molecular motion and we shall not attempt to derive it. Basically, this property states that in equilibrium thermal energy tends to distribute itself as evenly as possible consistent with the boundary conditions. It can be shown that the mean-value property uniquely determines the equilibrium temperature distribution of a plate.

Unfortunately, the determination of the equilibrium temperature distribution from the mean-value property is not an easy matter. However, if we restrict ourselves to finding the temperature only at a finite set of points within the plate, the problem can be reduced to solving a linear system. We shall pursue this idea in the next section.

DISCRETE FORMULATION OF THE PROBLEM

Let us overlay our trapezoidal plate with a succession of finer and finer square nets or meshes (Fig. 7.3). In (a) we have a rather coarse net; in (b) we have a net with half the spacing as in (a); and in (c) we have a net with the spacing again reduced by half. The points of intersection of the net lines are called *mesh points*. We classify them as *boundary mesh points* if they fall on the boundary of the plate or *interior mesh points* if they lie in the interior of the plate. For the three net spacings we have chosen, there are 1, 9, and 49 interior mesh points respectively.

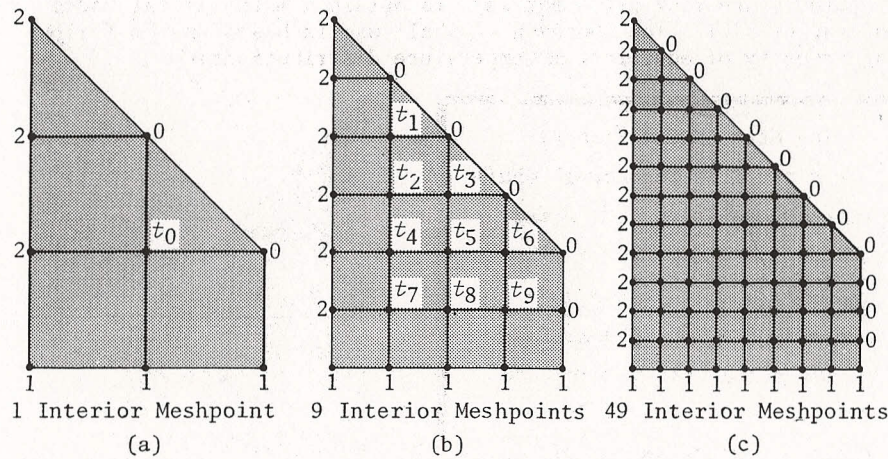


Figure 7.3

In the discrete formulation of our problem, we try to find the temperature only at the interior mesh points of some particular net. For a rather fine net, such as in (c), this will provide an excellent picture of the temperature distribution throughout the entire plate.

At the boundary mesh points, the temperature is given by the boundary data. (In Fig. 7.3 we have labeled all of the boundary mesh points with their corresponding temperatures.) At the interior mesh points, we shall apply the following discrete version of the mean - value property:

Discrete Mean - Value Property

At each interior mesh point, the temperature is the average of the temperatures at the four neighboring mesh points.

This discrete version is a reasonable approximation to the true mean - value property. But since it is only an approximation, it will provide only an approximation to the true temperatures at the interior mesh points. However, the approximations will get better as the mesh spacing decreases. In fact, as the mesh spacing approaches zero, the approximations approach the exact temperature distribution, a fact proved in advanced courses in Numerical

Analysis. We shall illustrate this convergence by computing the approximate temperatures at the mesh points for the three mesh spacings given in Fig. 7.3.

Case (a) of Fig. 7.3 is quite simple since there is only one interior mesh point. If we let t_0 be the temperature at this mesh point, the discrete mean - value property immediately gives

$$t_0 = \frac{1}{4} (2 + 1 + 0 + 0) = .75.$$

In case (b), let us label the temperatures at the nine interior mesh points t_1, t_2, \dots, t_9 , as in Fig. 7.3(b). (The particular ordering is not important.) By applying the discrete mean - value property successively to each of these nine mesh points, we obtain the following nine equations:

$$\begin{aligned} t_1 &= \frac{1}{4} (t_2 + 2 + 0 + 0) \\ t_2 &= \frac{1}{4} (t_1 + t_3 + t_4 + 2) \\ t_3 &= \frac{1}{4} (t_2 + t_5 + 0 + 0) \\ t_4 &= \frac{1}{4} (t_2 + t_5 + t_7 + 2) \\ t_5 &= \frac{1}{4} (t_3 + t_4 + t_6 + t_8) \\ t_6 &= \frac{1}{4} (t_5 + t_9 + 0 + 0) \\ t_7 &= \frac{1}{4} (t_4 + t_8 + 1 + 2) \\ t_8 &= \frac{1}{4} (t_5 + t_7 + t_9 + 1) \\ t_9 &= \frac{1}{4} (t_6 + t_8 + 1 + 0) \end{aligned} \tag{7.1}$$

This is a linear system of nine equations in nine unknowns. We may rewrite it in matrix form as

$$t = Mt + b \tag{7.2}$$

where

$$t = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \end{bmatrix}, \quad M = \begin{bmatrix} 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \\ 1/2 \\ 0 \\ 0 \\ 3/4 \\ 1/4 \\ 1/4 \end{bmatrix}.$$

To solve Eq. (7.2), let us write it as

$$(I - M)t = b.$$

The solution for t is thus

$$t = (I - M)^{-1}b \quad (7.3)$$

as long as the matrix $(I - M)$ is invertible. This is indeed the case, and the solution for t as calculated by (7.3) is

$$t = \begin{bmatrix} 0.7846 \\ 1.1383 \\ 0.4719 \\ 1.2967 \\ 0.7491 \\ 0.3265 \\ 1.2995 \\ 0.9014 \\ 0.5570 \end{bmatrix} \quad (7.4)$$

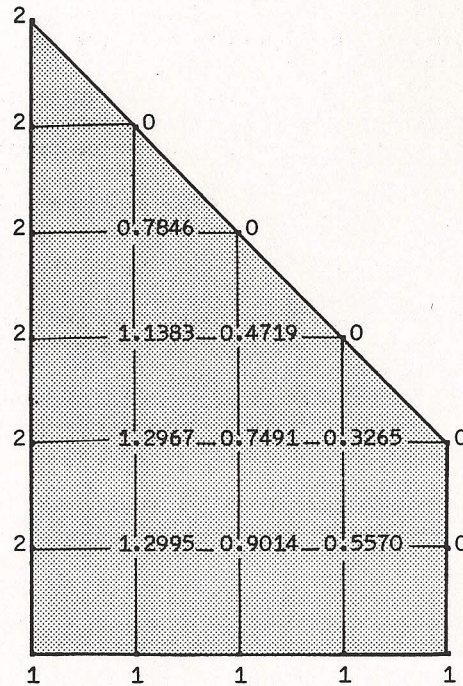


Figure 7.4

Figure 7.4 is a diagram of the plate with the nine interior mesh points labeled with their temperatures as given by this solution.

For case (c) of Fig. 7.3, we repeat this same procedure. We label the temperatures at the 49 interior mesh points as t_1, t_2, \dots, t_{49} in some manner. For example, we may begin at the top of the plate and proceed from left to right along each row of mesh points. Applying the discrete mean-value property to each mesh point gives a linear system of 49 equations in 49 unknowns:

$$\begin{aligned} t_1 &= \frac{1}{4}(t_2 + 2 + 0 + 0) \\ t_2 &= \frac{1}{4}(t_1 + t_3 + t_4 + 2) \\ &\vdots \\ t_{48} &= \frac{1}{4}(t_{41} + t_{47} + t_{49} + 1) \\ t_{49} &= \frac{1}{4}(t_{42} + t_{48} + 0 + 1). \end{aligned} \quad (7.5)$$

In matrix form, Eqs. (7.5) are

$$t = Mt + b$$

where t and b are column vectors with 49 entries and M is a 49×49 matrix. As in (7.3), the solution for t is

$$t = (I - M)^{-1}b. \quad (7.6)$$

In Fig. 7.5 we display the temperatures at the 49 mesh points found by Eq. (7.6). The nine unshaded temperatures in this figure fall on the same mesh points of Fig. 7.4. In Table 7.1 we compare the temperatures at these nine common mesh points for the three different mesh spacings used.

Temperatures at Common Mesh Points			
	Case (a)	Case (b)	Case (c)
t_1	—	0.7846	0.8048
t_2	—	1.1383	1.1533
t_3	—	0.4719	0.4778
t_4	—	1.2967	1.3078
t_5	0.7500	0.7491	0.7513
t_6	—	0.3265	0.3157
t_7	—	1.2995	1.3042
t_8	—	0.9014	0.9032
t_9	—	0.5570	0.5554

Table 7.1

Knowing that the temperatures of the discrete problem approach the exact temperatures as the mesh spacing decreases, we may surmise that the 9 temperatures obtained in case (c) are closer to the exact values than those in case (b).

A NUMERICAL TECHNIQUE

To obtain the 49 temperatures in case (c) above, it was necessary to solve a linear system with 49 unknowns. A finer net might involve a linear system with hundreds or even thousands of unknowns. Exact algorithms for the solutions of such large systems are impractical, and for this reason we shall now discuss a numerical technique for the practical solution of these systems.

To describe this technique, let us look again at Eq. 7.2:

$$t = Mt + b. \tag{7.7}$$

The vector t we are seeking appears on both sides of this equation. We consider a way of generating better and better approximations to the vector solution t . For the initial approximation $t^{(0)}$ we may take $t^{(0)} = 0$ if no better choice is available. If we substitute $t^{(0)}$ into the righthand side of (7.7) and label the resulting left-hand side as $t^{(1)}$, we have

$$t^{(1)} = Mt^{(0)} + b. \tag{7.8}$$

Usually $t^{(1)}$ is a better approximation to the solution than is $t^{(0)}$. If we substitute $t^{(1)}$ into the righthand side of (7.7) we generate another approximation, which we label as $t^{(2)}$:

$$t^{(2)} = Mt^{(1)} + b. \tag{7.9}$$

Continuing in this way, we generate a sequence of approximations as follows:

$$\begin{aligned} t^{(1)} &= Mt^{(0)} + b \\ t^{(2)} &= Mt^{(1)} + b \\ t^{(3)} &= Mt^{(2)} + b \\ &\vdots \\ t^{(n)} &= Mt^{(n-1)} + b \\ &\vdots \end{aligned} \tag{7.10}$$

Hopefully, this sequence of approximations $t^{(0)}, t^{(1)}, t^{(2)}, \dots$ converges to the exact solution of (7.7). We do not have the space in this book to go into the theoretical considerations necessary to show this. However, suffice it to say that for the particular

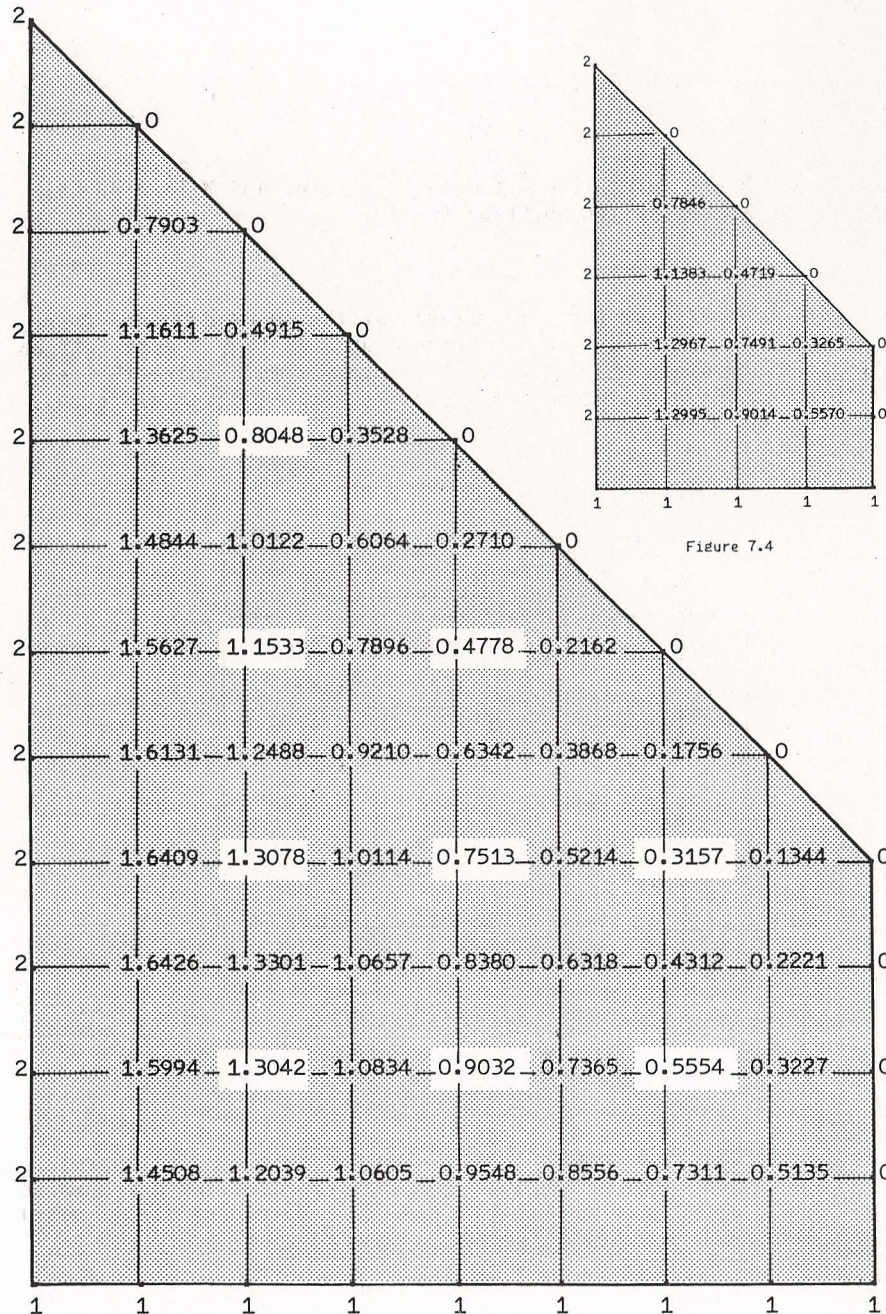


Figure 7.5

problem we are considering, the sequence converges to the exact solution for any mesh size and for any initial approximation $t^{(0)}$.

This technique of generating successive approximations to the solution of a linear system is called *Jacobi iteration*, and the approximations themselves are called *iterates*. As a numerical example, let us apply Jacobi iteration to the calculation of the nine mesh point temperatures of case (b). Setting $t^{(0)} = 0$, we have from Eq. (7.2)

$$t^{(1)} = Mt^{(0)} + b = MO + b = b = \begin{bmatrix} .5000 \\ .5000 \\ .0000 \\ .5000 \\ .0000 \\ .7500 \\ .2500 \\ .2500 \end{bmatrix},$$

$$t^{(2)} = Mt^{(1)} + b$$

$$= \begin{bmatrix} 0 & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 \end{bmatrix} \begin{bmatrix} .5000 \\ .5000 \\ .0000 \\ .5000 \\ .0000 \\ .0000 \\ .7500 \\ .2500 \\ .2500 \end{bmatrix} + \begin{bmatrix} .5000 \\ .5000 \\ .0000 \\ .5000 \\ .0000 \\ .0000 \\ .7500 \\ .2500 \\ .2500 \end{bmatrix} = \begin{bmatrix} .6250 \\ .7500 \\ .1250 \\ .8125 \\ .1875 \\ .0625 \\ .9375 \\ .5000 \\ .3125 \end{bmatrix}.$$

Some additional iterates are

$$t^{(3)} = \begin{bmatrix} 0.6875 \\ 0.8906 \\ 0.2344 \\ 0.9688 \\ 0.3750 \\ 0.1250 \\ 1.0781 \\ 0.6094 \\ 0.3906 \end{bmatrix}, \quad t^{(10)} = \begin{bmatrix} 0.7791 \\ 1.1230 \\ 0.4573 \\ 1.2770 \\ 0.7236 \\ 0.3131 \\ 1.2848 \\ 0.8827 \\ 0.5446 \end{bmatrix}, \quad t^{(20)} = \begin{bmatrix} 0.7845 \\ 1.1380 \\ 0.4716 \\ 1.2963 \\ 0.7486 \\ 0.3263 \\ 1.2992 \\ 0.9010 \\ 0.5567 \end{bmatrix}, \quad t^{(30)} = \begin{bmatrix} 0.7846 \\ 1.1383 \\ 0.4719 \\ 1.2967 \\ 0.7491 \\ 0.3265 \\ 1.2995 \\ 0.9014 \\ 0.5570 \end{bmatrix}.$$

All iterates beginning with the thirtieth are equal to $t^{(30)}$ to four decimal places. Consequently, $t^{(30)}$ is the exact solution to four decimal places. This agrees with our previous result given in Eq. (7.4).

The Jacobi iteration scheme applied to the linear system (7.5) with 49 unknowns produced iterates which begin repeating to four decimal places after 119 iterations. Thus, $t^{(119)}$ would provide the 49 temperatures of case (c) correct to four decimal places.

A MONTE CARLO TECHNIQUE

In this section, we describe a so-called *Monte Carlo technique* for computing the temperature at a single interior mesh point of the discrete problem without having to compute the temperatures at the remaining interior mesh points. First we define a *discrete random walk along the net*. By this we mean a directed path along the net lines (Fig. 7.6) which joins a succession of mesh points such that the direction of departure from each mesh point is chosen at random. Each of the four possible directions of departure from each mesh point along the path is to be equally probable.

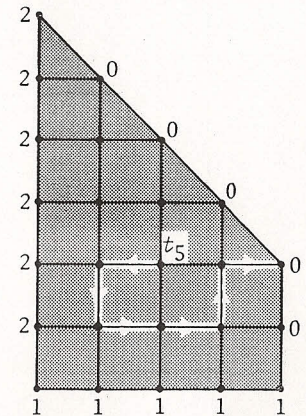


Figure 7.6

By the use of random walks, we can compute the temperature at a specified interior mesh point on the basis of the following property:

Random Walk Property

Let W_1, W_2, \dots, W_n be a succession of random walks, all of which begin at a specified interior mesh point. Let $t_1^*, t_2^*, \dots, t_n^*$ be the temperatures at the boundary mesh points first encountered along each of these random walks. Then the average value $(t_1^* + t_2^* + \dots + t_n^*)/n$ of these boundary temperatures approaches the temperature at the specified interior mesh point as the number of random walks n increases without bound.

This property is a consequence of the discrete mean-value property which the mesh point temperatures satisfy. The proof of the random walk property involves elementary concepts from probability theory and we shall not give it here.

In Table 7.2 we display the results of a large number of computer-generated random walks for the evaluation of the temperature t_5 of the nine-point mesh of case (b) (Fig. 7.6). The first column lists the number n of the random walk. The second column lists the temperature t_n^* of the boundary point first encountered along the corresponding random walk. The last column contains the cumulative average of the boundary temperatures encountered along the n random walks.

n	t_n^*	$(t_1^* + \dots + t_n^*)/n$	n	t_n^*	$(t_1^* + \dots + t_n^*)/n$
1	1	1.0000	20	1	0.9500
2	2	1.5000	30	0	0.8000
3	1	1.3333	40	0	0.8250
4	0	1.0000	50	2	0.8400
5	2	1.2000	100	0	0.8300
6	0	1.0000	150	1	0.8000
7	2	1.1429	200	0	0.8050
8	0	1.0000	250	1	0.8240
9	2	1.1111	500	1	0.7860
10	0	1.0000	1000	0	0.7550

Table 7.2

Thus, after 1000 random walks we have the approximation $t_5 \approx .7550$. This compares with the exact value $t_5 = .7491$ we had previously evaluated. As can be seen, the convergence to the exact value is not too rapid.

EXERCISES

7.1 A plate in the form of a circular disk has boundary temperatures of 0° on half of its circumference and 1° on the remaining half of its circumference. A net with four interior mesh points is overlaid on the disk (Fig. 7.7).

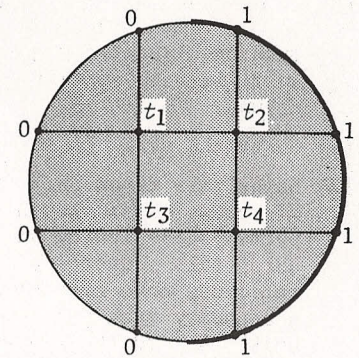


Figure 7.7

(a) Using the discrete mean-value property, write the 4×4 linear system $t = Mt + b$ which determines the approximate temperatures at the four interior mesh points.

(b) Solve the linear system in part (a).

(c) Use the Jacobi iteration scheme with $t^{(0)} = 0$ to generate the iterates $t^{(1)}, t^{(2)}, t^{(3)}, t^{(4)},$ and $t^{(5)}$ for the linear system in part (a). What is the "error vector" $t^{(5)} - t$ where t is the solution found in part (b)?

(d) By certain advanced methods, it can be determined that the exact temperatures at the four mesh points are $t_1 = t_3 = .2371$ and $t_2 = t_4 = .7629$. What are the percentage errors in the values found in part (b)?

7.2 Using the exact mean-value property (page 81) find the exact equilibrium temperature at the center of the disk in Exercise 7.1.

7.3 Calculate the first two iterates $t^{(1)}$ and $t^{(2)}$ for case (b) with nine interior mesh points (Eq. (7.2)) when the initial iterate is chosen as

$$t^{(0)} = [1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1 \ 1] t.$$

7.4 The random walk illustrated in Fig. 7.6 may be described by six arrows

← ↓ → → ↑ →

which specify the directions of departure from the successive mesh points along the path. Below is an array of 100 computer-generated, randomly-oriented arrows arranged in a 10×10 array. Use these arrows to determine ten random walks to approximate the temperature t_5 as in Table 7.2. Proceed as follows:

- (1) Take the last two digits of your telephone number. Use the last digit to specify a row and the other to specify a column.
- (2) Go to the arrow in the array with that row and column number.
- (3) Using this arrow as a starting point, move through the array of arrows as you would read a book (left to right and top to bottom). Beginning at the point labeled t_5 in Fig. 7.6 and using this sequence of arrows to specify a sequence of directions, move from mesh point to mesh point until you reach a boundary mesh point. This completes your first random walk. Record the temperature at the boundary mesh point. (If you reach the end of the arrow array, continue with the arrow in the upper lefthand corner.)
- (4) Return to the interior mesh point labeled t_5 and begin where you left off in the arrow array and generate your next random walk. Repeat this process until you have completed ten random walks and have recorded ten boundary temperatures.
- (5) Calculate the average of the ten boundary temperatures recorded. (The exact value is $t_5 = .7491$.)

	0	1	2	3	4	5	6	7	8	9
0	↓	←	←	↓	↓	←	←	↓	↓	↑
1	←	←	→	→	←	↑	↑	→	→	↓
2	←	↑	←	←	←	→	←	↓	←	↓
3	↑	→	→	→	→	↑	↑	←	↓	↑
4	↑	↑	←	←	↓	↓	↓	→	←	↓
5	↓	←	→	↓	↓	↑	↑	←	↑	↓
6	↑	↑	↑	↓	↓	↓	↑	↑	←	←
7	←	↓	↓	↑	→	→	↑	↓	↓	←
8	←	↑	→	↓	←	↓	↓	←	↓	←
9	↑	↑	→	→	↑	↓	↓	→	↓	→

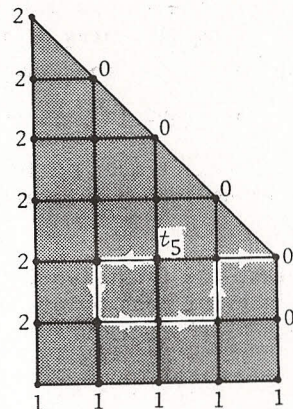


Figure 7.6

Some Applications in Genetics

Using the diagonalization of a matrix to compute its powers, the propagation of an inherited trait in successive generations is investigated.

PREREQUISITES: Eigenvalues and eigenvectors
 Diagonalization of a matrix
 Intuitive understanding of limits

INTRODUCTION

In this chapter, we shall examine the inheritance of traits in animals or plants. The inherited trait under consideration is assumed to be governed by a set of two genes, which we designate by the letters A and a . Under *autosomal inheritance* each individual in the population, of either sex, possesses two of these genes: the possible pairings being designated AA , Aa , and aa . This pair of genes is called the individual's *genotype*, and it determines how the trait controlled by the genes is manifested in the individual. For example, in snapdragons a set of two genes determines the color of the flower. Genotype AA produces red flowers, genotype Aa produces pink flowers, and genotype aa produces white flowers. In humans, eye coloration is controlled through autosomal inheritance. Genotypes AA and Aa have brown eyes, and genotype aa has blue eyes. In

this case, we say that gene A *dominates* gene a , or that gene a is *recessive* to gene A , since genotype Aa has the same outward trait as genotype AA .

In addition to autosomal inheritance, we shall also discuss *X-linked inheritance*. In this type of inheritance the male of the species possesses only one of the two possible genes (A or a), and the female possesses a pair of the two genes (AA , Aa , or aa). In humans, color blindness, hereditary baldness, hemophilia, and muscular dystrophy, to name a few, are traits controlled by X-linked inheritance.

In the next sections, we explain the manner in which the genes of the parents are passed on to their offspring for the two types of inheritance. We construct matrix models which give the probable genotypes of the offspring in terms of the genotypes of the parents, and use these matrix models to follow the genotype distribution of a population through successive generations.

AUTOSOMAL INHERITANCE

In autosomal inheritance, an individual inherits one gene from each of its parents' pairs of genes to form its own particular pair. As far as we know, it is a matter of chance which of the two genes a parent passes on to the offspring. Thus, if one parent is of genotype Aa , it is equally likely that the offspring will inherit the A gene or the a gene from that parent. If one parent is of genotype aa and the other parent is of genotype Aa , the offspring will always receive an a gene from the aa parent, and will receive either an A gene or an a gene, with equal probability, from the Aa parent. Consequently, the offspring has equal probability of being genotype AA or Aa . In Table 8.1, we list the probabilities of the possible genotypes of the offspring for all possible combinations of the genotypes of the parents.

		Genotypes of Parents					
		$AA - AA$	$AA - Aa$	$AA - aa$	$Aa - Aa$	$Aa - aa$	$aa - aa$
Genotype of Offspring	AA	1	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0
	Aa	0	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$	0
	aa	0	0	0	$\frac{1}{4}$	$\frac{1}{2}$	1

Table 8.1

EXAMPLE 8.1 Suppose a farmer has a large population of plants consisting of some distribution of all three possible genotypes AA , Aa , and aa . He desires to undertake a breeding program in which each plant in the population is always fertilized with a plant of genotype AA . We want to derive an expression for the distribution of the three possible genotypes in the population after any number of generations.

For $n = 0, 1, 2, \dots$, let us set

a_n = fraction of plants of genotype AA in n -th generation,

b_n = fraction of plants of genotype Aa in n -th generation,

c_n = fraction of plants of genotype aa in n -th generation.

Thus, a_0 , b_0 , and c_0 specify the initial distribution of the genotypes. We also have that

$$a_n + b_n + c_n = 1 \quad \text{for } n = 0, 1, 2, \dots$$

From Table 8.1, we may determine the genotype distribution of each generation from the genotype distribution of the preceding generation by the following equations:

$$a_n = a_{n-1} + \frac{1}{2}b_{n-1}$$

$$b_n = c_{n-1} + \frac{1}{2}b_{n-1} \quad n = 1, 2, \dots \quad (8.1)$$

$$c_n = 0.$$

For example, the first of these three equations states that all the offspring of a plant of genotype AA will be of genotype AA under this breeding program, and half of the offspring of a plant of genotype Aa will be of genotype AA .

Equations (8.1) can be written in matrix notation as

$$\mathbf{x}^{(n)} = M\mathbf{x}^{(n-1)} \quad n = 1, 2, \dots \quad (8.2)$$

where

$$\mathbf{x}^{(n)} = \begin{bmatrix} a_n \\ b_n \\ c_n \end{bmatrix}, \quad \mathbf{x}^{(n-1)} = \begin{bmatrix} a_{n-1} \\ b_{n-1} \\ c_{n-1} \end{bmatrix}, \quad \text{and} \quad M = \begin{bmatrix} 1 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Notice that the three columns of the matrix M are the same as the first three columns of Table 8.1.

From Eq. (8.2) it follows that

$$\mathbf{x}^{(n)} = M\mathbf{x}^{(n-1)} = M^2\mathbf{x}^{(n-2)} = \dots = M^n\mathbf{x}^{(0)}. \quad (8.3)$$

Consequently, if we can find an explicit expression for M^n , we can use Eq. (8.3) to obtain an explicit expression $\mathbf{x}^{(n)}$. To find an explicit expression for M^n , we first diagonalize M . That is, we find an invertible matrix P and a diagonal matrix D such that

$$M = PDP^{-1}. \quad (8.4)$$

With such a diagonalization, we then have (see Exercise 8.1)

$$M^n = PD^nP^{-1} \quad \text{for } n = 1, 2, \dots$$

where

$$D^n = \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_k \end{bmatrix}^n = \begin{bmatrix} d_1^n & 0 & 0 & \dots & 0 \\ 0 & d_2^n & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_k^n \end{bmatrix}.$$

The diagonalization of M is accomplished by finding its eigenvalues and corresponding eigenvectors. These are easily found to be (verify)

$$\text{Eigenvalues: } \lambda_1 = 1, \quad \lambda_2 = \frac{1}{2}, \quad \lambda_3 = 0$$

$$\text{Corresponding Eigenvectors: } \mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

In Eq. (8.4), we then have

$$D = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$P = \left[\begin{array}{c|c|c} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \end{array} \right] = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, we have

$$\mathbf{x}^{(n)} = PD^nP^{-1}\mathbf{x}^{(0)} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & (\frac{1}{2})^n & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \\ c_0 \end{bmatrix}$$

or

$$\mathbf{x}^{(n)} = \begin{bmatrix} a_n \\ b_n \\ c_n \end{bmatrix} = \begin{bmatrix} 1 & 1 - (\frac{1}{2})^n & 1 - (\frac{1}{2})^{n-1} \\ 0 & (\frac{1}{2})^n & (\frac{1}{2})^{n-1} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \\ c_0 \end{bmatrix} \\ = \begin{bmatrix} a_0 + b_0 + c_0 - (\frac{1}{2})^n b_0 - (\frac{1}{2})^{n-1} c_0 \\ (\frac{1}{2})^n b_0 + (\frac{1}{2})^{n-1} c_0 \\ 0 \end{bmatrix}.$$

Using the fact that $a_0 + b_0 + c_0 = 1$, we thus have

$$\begin{aligned} a_n &= 1 - (\frac{1}{2})^n b_0 - (\frac{1}{2})^{n-1} c_0 \\ b_n &= (\frac{1}{2})^n b_0 + (\frac{1}{2})^{n-1} c_0 \quad n = 1, 2, \dots \quad (8.5) \\ c_n &= 0. \end{aligned}$$

These are explicit formulas for the fractions of the three genotypes in the n -th generation of plants in terms of the initial genotype fractions.

Since $(\frac{1}{2})^n$ tends to zero as n approaches infinity, it follows from these equations that

$$\begin{aligned} a_n &\rightarrow 1 \\ b_n &\rightarrow 0 \\ c_n &= 0 \end{aligned}$$

as n approaches infinity. That is, in the limit all plants in the population will be genotype AA .

EXAMPLE 8.2 Let us modify Example 8.1 so that instead of fertilizing each plant with one of genotype AA , each plant is fertilized with a plant of its own genotype. Using the same notation as in Example 8.1, we then find

$$\mathbf{x}^{(n)} = M^n \mathbf{x}^{(0)}$$

where

$$M = \begin{bmatrix} 1 & \frac{1}{4} & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & 1 \end{bmatrix}.$$

The columns of this new matrix M are the same as the columns of Table 8.1 corresponding to parents with genotypes $AA-AA$, $Aa-Aa$, and $aa-aa$.

The eigenvalues of M are easily found to be (verify)

$$\lambda_1 = 1, \quad \lambda_2 = 1, \quad \lambda_3 = \frac{1}{2}.$$

The eigenvalue $\lambda_1 = 1$ has multiplicity two and its corresponding eigenspace is two-dimensional. Picking two linearly independent eigenvectors \mathbf{e}_1 and \mathbf{e}_2 in that eigenspace, and a single eigenvector \mathbf{e}_3 for the simple eigenvalue $\lambda_3 = \frac{1}{2}$, we have (verify)

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}.$$

The calculations for $\mathbf{x}^{(n)}$ are then

$$\begin{aligned} \mathbf{x}^{(n)} &= M^n \mathbf{x}^{(0)} = P D^n P^{-1} \mathbf{x}^{(0)} \\ &= \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & -2 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & (\frac{1}{2})^n \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & 0 \\ 1 & 1 & 1 \\ 0 & -\frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \\ c_0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & \frac{1}{2} - (\frac{1}{2})^{n+1} & 0 \\ 0 & (\frac{1}{2})^n & 0 \\ 0 & \frac{1}{2} - (\frac{1}{2})^{n+1} & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \\ c_0 \end{bmatrix}. \end{aligned}$$

We thus have

$$\begin{aligned} a_n &= a_0 + [\frac{1}{2} - (\frac{1}{2})^{n+1}] b_0 \\ b_n &= (\frac{1}{2})^n b_0 \\ c_n &= c_0 + [\frac{1}{2} - (\frac{1}{2})^{n+1}] b_0. \end{aligned} \quad n = 1, 2, \dots \quad (8.6)$$

In the limit, as n tends to infinity, $(\frac{1}{2})^n \rightarrow 0$ and $(\frac{1}{2})^{n+1} \rightarrow 0$ so that

$$\begin{aligned} a_n &\rightarrow a_0 + \frac{1}{2} b_0 \\ b_n &\rightarrow 0 \\ c_n &\rightarrow c_0 + \frac{1}{2} b_0. \end{aligned}$$

Thus, fertilization of each plant with one of its own genotype produces a population which in the limit contains only genotypes AA and aa .

AUTOSOMAL RECESSIVE DISEASES

In human populations, there are many genetic diseases governed by autosomal inheritance in which a normal gene A dominates an abnormal gene a . Genotype AA is a normal individual; genotype Aa is a carrier of the disease, though is not afflicted with the disease; and genotype aa is afflicted with the disease. In some cases, such genetic diseases are associated with a particular racial group; e.g., cystic fibrosis (predominant among Caucasians), sickle cell anemia (predominant among Blacks), Cooley's anemia (predominant among people of Mediterranean origin), and Tay-Sachs disease (predominant among Eastern European Jews). It is quite often the case with such diseases that the afflicted individuals die before reaching maturity, so that all afflicted children are the offspring of parents who are both carriers. Suppose a program to identify the carriers of such a disease is carried out, and all of the carriers identified agree not to produce offspring among themselves. In this way, all future children will either have two normal parents ($AA-AA$ matings) or one normal parent and one carrier parent ($AA-Aa$ matings). Consequently, no future children will be afflicted with the disease, though there will still be carriers in future generations. Let us now determine the fraction of carriers in future generations under such a controlled mating program. We set

$$\mathbf{x}^{(n)} = \begin{bmatrix} a_n \\ b_n \end{bmatrix}, \quad n = 1, 2, \dots$$

where

α_n = fraction of population of genotype AA in n -th generation,

b_n = fraction of population of genotype Aa (carriers) in n -th generation.

Since each offspring has at least one normal parent, we may consider the controlled mating program as one of continual mating with genotype AA , as in Example 8.1. Thus, the transition of genotype distributions from one generation to the next is governed by the equation

$$\mathbf{x}^{(n)} = M\mathbf{x}^{(n-1)}, \quad n = 1, 2, \dots,$$

where

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

Knowing the initial distribution $\mathbf{x}^{(0)}$, the distribution of genotypes in the n -th generation is thus given by

$$\mathbf{x}^{(n)} = M^n \mathbf{x}^{(0)}, \quad n = 1, 2, \dots$$

The diagonalization of M is easily carried out (see Exercise 8.4), and leads to

$$\begin{aligned} \mathbf{x}^{(n)} &= PD^n P^{-1} \mathbf{x}^{(0)} = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (\frac{1}{2})^n \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ b_0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 - (\frac{1}{2})^n \\ 0 & (\frac{1}{2})^n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ b_0 \end{bmatrix} = \begin{bmatrix} \alpha_0 + b_0 - (\frac{1}{2})^n b_0 \\ (\frac{1}{2})^n b_0 \end{bmatrix}. \end{aligned}$$

Since $\alpha_0 + b_0 = 1$, we have

$$\begin{aligned} \alpha_n &= 1 - (\frac{1}{2})^n b_0 \\ b_n &= (\frac{1}{2})^n b_0. \end{aligned} \quad n = 1, 2, \dots \quad (8.7)$$

As n tends to infinity, we thus have

$$\alpha_n \rightarrow 1$$

$$b_n \rightarrow 0,$$

so that in the limit there will be no carriers in the population. From (8.7) we see that

$$b_n = \frac{1}{2} b_{n-1}, \quad n = 1, 2, \dots \quad (8.8)$$

That is, the fraction of carriers in each generation is one-half the fraction of carriers in the preceding generation. It would be of interest to also investigate the propagation of carriers under random mating, when two individuals mate without regard to their genotypes. Unfortunately, such random mating leads to nonlinear equations, and the techniques of this chapter are not applicable. However, by other techniques it can be shown that under random mating Eq. (8.8) is replaced by

$$b_n = \frac{b_{n-1}}{1 + \frac{1}{2} b_{n-1}}, \quad n = 1, 2, \dots \quad (8.9)$$

As a numerical example, in the United States, approximately 10% of Blacks are presently carriers of the gene for sickle cell anemia. Under a controlled mating program governed by Eq. (8.8), the percentage of carriers can be reduced to 5% in one generation (about 27 years). But under random mating, Eq. (8.9) predicts that 9.5% of the population will be carriers after one generation (i.e., $b_n = .095$ if $b_{n-1} = .10$). In addition, under controlled mating no offspring will ever be afflicted with the disease, but with random mating it can be shown that about 1 in 400 Black children will be born with sickle cell anemia when 10% of the population are carriers.

X-LINKED INHERITANCE

As mentioned in the introduction, in X-linked inheritance the male possesses one gene (A or a), and the female possesses two genes (AA , Aa , or aa). The term "X-linked" is used because such genes are found on the X-chromosome, of which the male has one and the female has two. The inheritance of such genes is as follows: a male offspring receives one of his mothers two genes with equal probability, and a female offspring receives the one gene of her father and one of her mother's two genes with equal probability. Readers familiar with basic probability can verify that this type of inheritance leads to the following table of genotype probabilities.

		Genotypes of Parents (father, mother)						
		(A, AA)	(A, Aa)	(A, aa)	(a, AA)	(a, Aa)	(a, aa)	
Genotype of Offspring	Male	A	1	1/2	0	1	1/2	0
		a	0	1/2	1	0	1/2	1
Female	AA	1	1/2	0	0	0	0	0
	Aa	0	1/2	1	1	1/2	0	0
	aa	0	0	0	0	1/2	1	0

Table 8.2

We shall discuss a program of inbreeding in connection with X-linked inheritance. We begin initially with a male and female; select two of their offspring at random, one of each sex, and mate them; select two of the resulting offspring and mate them; and so forth. Such inbreeding is commonly performed with animals. Among humans, such brother-sister marriages were used by the rulers of ancient Egypt to keep the royal line pure.

The original male-female pair can be one of the six types, corresponding to the six columns of Table 8.2:

$$(A, AA), (A, Aa), (A, aa), (a, AA), (a, Aa), (a, aa).$$

The sibling-pairs mated in each successive generation have certain probabilities of being one of these six types. To compute these probabilities, for $n = 0, 1, 2, \dots$, let us set

a_n	=	probability sibling-pair mated in n -th generation is type (A, AA)
b_n	=	" " " " " " " " (A, Aa)
c_n	=	" " " " " " " " (A, aa)
d_n	=	" " " " " " " " (a, AA)
e_n	=	" " " " " " " " (a, Aa)
f_n	=	" " " " " " " " (a, aa)

With these probabilities we form a column vector

$$x^{(n)} = \begin{bmatrix} a_n \\ b_n \\ c_n \\ d_n \\ e_n \\ f_n \end{bmatrix} \quad n = 0, 1, 2, \dots$$

From Table 8.2, it is found that

$$x^{(n)} = Mx^{(n-1)}, \quad n = 1, 2, \dots \quad (8.10)$$

where

$$M = \begin{matrix} & \begin{matrix} (A,AA) & (A,Aa) & (A,aa) & (a,AA) & (a,Aa) & (a,aa) \end{matrix} \\ \begin{matrix} (A,AA) \\ (A,Aa) \\ (A,aa) \\ (a,AA) \\ (a,Aa) \\ (a,aa) \end{matrix} & \begin{bmatrix} 1 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 1 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 0 \\ 0 & 1/4 & 0 & 0 & 0 & 0 \\ 0 & 1/4 & 1 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 1/4 & 1 \end{bmatrix} \end{matrix}$$

For example, suppose in the $(n-1)$ st generation the sibling-pair mated is type (A, Aa) . Then their male offspring will be either genotype A or a with equal probability, and their female offspring will be either genotype AA or Aa with equal probability. Since one of the male offspring and one of the female offspring are chosen at random for mating, the next sibling-pair will be one of type (A, AA) , (A, Aa) , (a, AA) , (a, Aa) , with equal probability. Thus, the second column of M contains "1/4" in each of the four rows corresponding to these four sibling-pairs. (See Exercise 8.9 for the remaining columns.)

As in our previous examples, from (8.10) we obtain

$$x^{(n)} = M^n x^{(0)}, \quad n = 1, 2, \dots \quad (8.11)$$

After some lengthy calculations, the eigenvalues and eigenvectors of M turn out to be

$$\lambda_1 = 1, \quad \lambda_2 = 1, \quad \lambda_3 = 1/2, \quad \lambda_4 = -1/2, \quad \lambda_5 = 1/4(1 + \sqrt{5}), \quad \lambda_6 = 1/4(1 - \sqrt{5})$$

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad e_3 = \begin{bmatrix} -1 \\ 2 \\ -1 \\ 1 \\ -2 \\ 1 \end{bmatrix}, \quad e_4 = \begin{bmatrix} 1 \\ -6 \\ -3 \\ 3 \\ 6 \\ -1 \end{bmatrix}, \quad e_5 = \begin{bmatrix} 1/4(-3 - \sqrt{5}) \\ 1 \\ 1/4(-1 + \sqrt{5}) \\ 1/4(-1 + \sqrt{5}) \\ 1 \\ 1/4(-3 - \sqrt{5}) \end{bmatrix}, \quad e_6 = \begin{bmatrix} 1/4(-3 + \sqrt{5}) \\ 1 \\ 1/4(-1 - \sqrt{5}) \\ 1/4(-1 - \sqrt{5}) \\ 1 \\ 1/4(-3 + \sqrt{5}) \end{bmatrix}$$

The diagonalization of M then leads to

$$\mathbf{x}^{(n)} = PD^nP^{-1}\mathbf{x}^{(0)}, \quad n = 1, 2, \dots \quad (8.12)$$

where

$$P = \begin{bmatrix} 1 & 0 & -1 & 1 & \frac{1}{4}(-3 - \sqrt{5}) & \frac{1}{4}(-3 + \sqrt{5}) \\ 0 & 0 & 2 & -6 & 1 & 1 \\ 0 & 0 & -1 & -3 & \frac{1}{4}(-1 + \sqrt{5}) & \frac{1}{4}(-1 - \sqrt{5}) \\ 0 & 0 & 1 & 3 & \frac{1}{4}(-1 + \sqrt{5}) & \frac{1}{4}(-1 - \sqrt{5}) \\ 0 & 0 & -2 & 6 & 1 & 1 \\ 0 & 1 & 1 & -1 & \frac{1}{4}(-3 - \sqrt{5}) & \frac{1}{4}(-3 + \sqrt{5}) \end{bmatrix}$$

$$D^n = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & (\frac{1}{2})^n & 0 & 0 & 0 \\ 0 & 0 & 0 & (-\frac{1}{2})^n & 0 & 0 \\ 0 & 0 & 0 & 0 & [\frac{1}{4}(1 + \sqrt{5})]^n & 0 \\ 0 & 0 & 0 & 0 & 0 & [\frac{1}{4}(1 - \sqrt{5})]^n \end{bmatrix}$$

$$P^{-1} = \begin{bmatrix} 1 & 2/3 & 1/3 & 2/3 & 1/3 & 0 \\ 0 & 1/3 & 2/3 & 1/3 & 2/3 & 1 \\ 0 & 1/8 & -1/4 & 1/4 & -1/8 & 0 \\ 0 & -1/24 & -1/12 & 1/12 & 1/24 & 0 \\ 0 & (5 + \sqrt{5})/20 & \sqrt{5}/5 & \sqrt{5}/5 & (5 + \sqrt{5})/20 & 0 \\ 0 & (5 - \sqrt{5})/20 & -\sqrt{5}/5 & -\sqrt{5}/5 & (5 - \sqrt{5})/20 & 0 \end{bmatrix}$$

We shall not write out the matrix product in (8.12) since it is rather unwieldy. However, if a specific vector $\mathbf{x}^{(0)}$ is given, the calculation for $\mathbf{x}^{(n)}$ is not too cumbersome (see Exercise 8.6).

Since the absolute values of the last four diagonal entries of D are less than one, we see that as n tends to infinity

$$D^n \rightarrow \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

And so from Eq. (8.12):

$$\mathbf{x}^{(n)} \rightarrow P \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} P^{-1} \mathbf{x}^{(0)}.$$

Performing the matrix multiplication on the right we obtain (verify)

$$\mathbf{x}^{(n)} \rightarrow \begin{bmatrix} a_0 + \frac{2}{3}b_0 + \frac{1}{3}c_0 + \frac{2}{3}d_0 + \frac{1}{3}e_0 \\ 0 \\ 0 \\ 0 \\ 0 \\ f_0 + \frac{1}{3}b_0 + \frac{2}{3}c_0 + \frac{1}{3}d_0 + \frac{2}{3}e_0 \end{bmatrix}. \quad (8.13)$$

That is, in the limit all sibling-pairs will be either type (A, AA) or type (a, aa) . For example, if the initial parents are type (A, Aa) (i.e., $b_0 = 1$ and $a_0 = c_0 = d_0 = e_0 = f_0 = 0$), then as n tends to infinity

$$\mathbf{x}^{(n)} \rightarrow \begin{bmatrix} 2/3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1/3 \end{bmatrix}$$

Thus, in the limit, there is probability $2/3$ that the sibling-pairs will all be (A, AA) , and probability $1/3$ that they will all be (a, aa) .

EXERCISES

- 8.1 Show that if $M = PDP^{-1}$, then $M^n = PD^nP^{-1}$ for $n = 1, 2, \dots$.
- 8.2 In Example 8.1, suppose the plants are always fertilized with a plant of genotype Aa , rather than one of genotype AA . Derive formulas for the fractions of the plants of genotypes AA , Aa , and aa in the n -th generation. Also, find the limiting genotype distribution as n tends to infinity.
- 8.3 In Example 8.1, suppose the initial plants are fertilized with genotype AA , the first generation is fertilized with genotype Aa , the second generation is fertilized with genotype AA , and this alternating pattern of fertilization is kept up. Find formulas for the fractions of the plants of genotypes AA , Aa , and aa in the n -th generation.
- 8.4 In the section on autosomal recessive diseases, find the eigenvalues and eigenvectors of the matrix M and verify Eqs. (8.7).
- 8.5 It is known that carriers of the sickle cell gene have a high resistance to malaria. Consequently, in regions where malaria is common, a balance is reached such that this benevolent aspect of the sickle cell gene balances its malevolent aspect of producing individuals with sickle cell anemia. In certain regions of Africa, this has resulted in an equilibrium situation in which up to 25% of the population are carriers. In the United States, however, malaria is not a problem. Thus, when Blacks from Africa were brought here, only the malevolent aspect of the sickle cell gene remained. The percentage of carriers then began to fall among the Blacks in this country, as described in this chapter, to the point where approximately 10% of Blacks are presently carriers. Using Eq. (8.9), calculate how many generations would be required for the percentage of carriers to fall from 25% to 10% when mating is irrespective of genotype. What would be the corresponding drop for the same number of generations under the controlled mating program determined by Eq. (8.8)?
- 8.6 In the section on X-linked inheritance, suppose the initial parents are equally likely to be of any of the six possible genotype parents; i.e.,

$$x^{(0)} = \begin{bmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \end{bmatrix}.$$

Using Eq. (8.12), calculate $x^{(n)}$, and also calculate the limit of $x^{(n)}$ as n tends to infinity.

- 8.7 From Eq. (8.13), show that under X-linked inheritance with inbreeding, the probability that the limiting sibling-pairs will be of type (A, AA) is the same as the proportion of A genes in the initial population.
- 8.8 In X-linked inheritance, suppose none of the females of genotype Aa survive to maturity. Under inbreeding the possible sibling-pairs are then
- $$(A, AA), (A, aa), (a, AA), \text{ and } (a, aa).$$

Find the transition matrix which describes how the genotype distribution changes in one generation.

- 8.9 Derive the matrix M in Eq. (8.10) from Table 8.2.

Age-Specific Population Growth

The growth in time of a female population divided into age classes is investigated using the Leslie matrix model. The limiting age distribution and growth rate of the population are determined.

PREREQUISITES: Eigenvalues and eigenvectors
Diagonalization of a matrix
Intuitive understanding of limits

INTRODUCTION

One of the most common models of population growth used by demographers is the so-called "Leslie model", developed in the nineteen forties. This model describes the growth of the female portion of a human or animal population. In this model, the females are divided into age classes of equal duration. To be specific, suppose the maximum age attained by any female in the population is L years (or some other time unit) and we divide the population into n age classes. Then each class is L/n years in duration. We label the age classes according to the following table:

Age Class	Age Interval
1	$[0, L/n]$
2	$[L/n, 2L/n]$
3	$[2L/n, 3L/n]$
\vdots	\vdots
$n-1$	$[(n-2)L/n, (n-1)L/n]$
n	$[(n-1)L/n, L]$

Suppose we know the number of females in each of the n classes at time $t=0$. In particular, let there be $x_1^{(0)}$ females in the first class, $x_2^{(0)}$ females in the second class, and so forth. With these n numbers we form a column vector $x^{(0)}$ as follows:

$$x^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix}$$

We call this vector the *initial age distribution vector*.

As time progresses, the number of females within each of the n classes changes because of three biological processes: birth, death, and aging. By describing these three processes quantitatively, we shall see how to project the initial age distribution vector into the future.

The easiest way to study the aging process is to observe the population at discrete times, say $t_0, t_1, t_2, \dots, t_k, \dots$. The Leslie model requires that the duration between any two successive observation times be the same as the duration of the age intervals. We therefore set

$$\begin{aligned} t_0 &= 0 \\ t_1 &= L/n \\ t_2 &= 2L/n \\ &\vdots \\ t_k &= kL/n \\ &\vdots \end{aligned}$$

With this assumption, all females in the $(i+1)$ st class at time t_{k+1} were in the i -th class at time t_k .

The birth and death processes between two successive observation times may be described by means of the following demographic parameters:

a_i $i = 1, 2, \dots, n$	The average number of daughters born to a single female during the time she is in the i -th age class.
b_i $i = 1, 2, \dots, n-1$	The fraction of females in the i -th age class that can be expected to survive and pass into the $(i+1)$ st age class.

By their definitions, we have that

- (i) $a_i \geq 0$ for $i = 1, 2, \dots, n$
- (ii) $0 < b_i \leq 1$ for $i = 1, 2, \dots, n-1$.

Notice that we do not allow any b_i to equal zero, since then no females will survive beyond the i -th age class. We also assume that at least one a_i is positive so that some births occur. Any age class for which the corresponding value of a_i is positive is called a *fertile age class*.

We next define the age distribution vector $x^{(k)}$ at time t_k by

$$x^{(k)} = \begin{bmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{bmatrix},$$

where $x_i^{(k)}$ is the number of females in the i -th age class at time t_k . Now, at time t_k , the females in the first age class are just those daughters born between times t_{k-1} and t_k . Thus we can write

$$\begin{pmatrix} \text{number of} \\ \text{females} \\ \text{in class 1} \\ \text{at time } t_k \end{pmatrix} = \begin{pmatrix} \text{number of} \\ \text{daughters} \\ \text{born to} \\ \text{females in} \\ \text{class 1} \\ \text{between times} \\ t_{k-1} \text{ and } t_k \end{pmatrix} + \begin{pmatrix} \text{number of} \\ \text{daughters} \\ \text{born to} \\ \text{females in} \\ \text{class 2} \\ \text{between times} \\ t_{k-1} \text{ and } t_k \end{pmatrix} + \dots + \begin{pmatrix} \text{number of} \\ \text{daughters} \\ \text{born to} \\ \text{females in} \\ \text{class } n \\ \text{between times} \\ t_{k-1} \text{ and } t_k \end{pmatrix}$$

or mathematically,

$$x_1^{(k)} = a_1 x_1^{(k-1)} + a_2 x_2^{(k-1)} + \dots + a_n x_n^{(k-1)} \quad (9.1)$$

The number of females in the $(i+1)$ st age class ($i = 1, 2, \dots, n-1$) at time t_k are those females in the i -th class at time t_{k-1} who are still alive at time t_k . Thus,

$$\begin{pmatrix} \text{number of} \\ \text{females in} \\ \text{class } i+1 \\ \text{at time } t_k \end{pmatrix} = \begin{pmatrix} \text{fraction of} \\ \text{females in} \\ \text{class } i \\ \text{who survive} \\ \text{and pass into} \\ \text{class } i+1 \end{pmatrix} \begin{pmatrix} \text{number of} \\ \text{females in} \\ \text{class } i \\ \text{at time } t_{k-1} \end{pmatrix}$$

or mathematically,

$$x_{i+1}^{(k)} = b_i x_i^{(k-1)}, \quad i = 1, 2, \dots, n-1 \quad (9.2)$$

Using matrix notation, Eqs. (9.1) and (9.2) can be written

$$\begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ x_3^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 & \dots & a_{n-1} & a_n \\ b_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & b_2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & b_{n-1} & 0 \end{pmatrix} \begin{pmatrix} x_1^{(k-1)} \\ x_2^{(k-1)} \\ x_3^{(k-1)} \\ \vdots \\ x_n^{(k-1)} \end{pmatrix},$$

or more compactly,

$$x^{(k)} = Lx^{(k-1)}, \quad k = 1, 2, \dots \quad (9.3)$$

where L is the *Leslie matrix*

$$L = \begin{pmatrix} a_1 & a_2 & a_3 & \dots & a_{n-1} & a_n \\ b_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & b_2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & b_{n-1} & 0 \end{pmatrix} \quad (9.4)$$

From Eq. (9.3) it follows that

$$\begin{aligned} x^{(1)} &= Lx^{(0)} \\ x^{(2)} &= Lx^{(1)} = L^2x^{(0)} \\ x^{(3)} &= Lx^{(2)} = L^3x^{(0)} \\ &\vdots \\ x^{(k)} &= Lx^{(k-1)} = L^kx^{(0)}. \end{aligned} \quad (9.5)$$

Thus, if we know the initial age distribution $x^{(0)}$ and the Leslie matrix L , we can determine the female age distribution at any later time.

EXAMPLE 9.1 Suppose the oldest age attained by the females in a certain animal population is 15 years, and we divide the population into three equal age classes of durations five years. Let the Leslie matrix for this population be

$$L = \begin{pmatrix} 0 & 4 & 3 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{pmatrix}$$

If there are initially 1,000 females in each of the three age classes, then from Eq. (9.3) we have

$$x^{(0)} = \begin{bmatrix} 1,000 \\ 1,000 \\ 1,000 \end{bmatrix},$$

$$x^{(1)} = Lx^{(0)} = \begin{bmatrix} 0 & 4 & 3 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{bmatrix} \begin{bmatrix} 1,000 \\ 1,000 \\ 1,000 \end{bmatrix} = \begin{bmatrix} 7,000 \\ 500 \\ 250 \end{bmatrix},$$

$$x^{(2)} = Lx^{(1)} = \begin{bmatrix} 0 & 4 & 3 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{bmatrix} \begin{bmatrix} 7,000 \\ 500 \\ 250 \end{bmatrix} = \begin{bmatrix} 2,750 \\ 3,500 \\ 125 \end{bmatrix},$$

$$x^{(3)} = Lx^{(2)} = \begin{bmatrix} 0 & 4 & 3 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{bmatrix} \begin{bmatrix} 2,750 \\ 3,500 \\ 125 \end{bmatrix} = \begin{bmatrix} 14,375 \\ 1,375 \\ 875 \end{bmatrix}.$$

Thus, after 15 years there are 14,375 females between 0 and 5 years of age, 1,375 females between 5 and 10 years of age, and 875 females between 10 and 15 years of age.

LIMITING BEHAVIOR

Although Eq. (9.5) gives the age distribution of the population at any time, it does not immediately give a general picture of the dynamics of the growth process. For this we need to investigate the eigenvalues and eigenvectors of the Leslie matrix. The eigenvalues of L are the roots of its characteristic polynomial. As we ask the reader to verify in Exercise 9.2, this characteristic polynomial is

$$p(\lambda) = |\lambda I - L|$$

$$= \lambda^n - a_1 \lambda^{n-1} - a_2 b_1 \lambda^{n-2} - a_3 b_1 b_2 \lambda^{n-3} - \dots - a_n b_1 b_2 \dots b_{n-1}.$$

To analyze the roots of this polynomial, it will be convenient to introduce the function

$$q(\lambda) = \frac{a_1}{\lambda} + \frac{a_2 b_1}{\lambda^2} + \frac{a_3 b_1 b_2}{\lambda^3} + \dots + \frac{a_n b_1 b_2 \dots b_{n-1}}{\lambda^n}. \quad (9.6)$$

Using this function, the characteristic equation $p(\lambda) = 0$ can be written (verify)

$$q(\lambda) = 1 \quad \text{for } \lambda \neq 0. \quad (9.7)$$

Since all of the a_i and b_i are nonnegative, we see that $q(\lambda)$ is monotonically decreasing for λ greater than zero. Furthermore, $q(\lambda)$ has a vertical asymptote at $\lambda = 0$ and approaches zero as $\lambda \rightarrow \infty$. Consequently, as Fig. 9.1 indicates, there is a unique λ , say λ_1 , such that $q(\lambda) = 1$. That is, the matrix L has a unique positive eigenvalue. It may further be shown (see Exercise 9.3) that λ_1 is simple; i.e., λ_1 has multiplicity one. Although we shall omit the computational details, the reader can verify that an eigenvector corresponding to λ_1 , that is, a nonzero vector solution of

$$Lx_1 = \lambda_1 x_1,$$

is

$$x_1 = \begin{bmatrix} 1 \\ b_1/\lambda_1 \\ b_1 b_2/\lambda_1^2 \\ b_1 b_2 b_3/\lambda_1^3 \\ \vdots \\ b_1 b_2 \dots b_{n-1}/\lambda_1^{n-1} \end{bmatrix}. \quad (9.8)$$

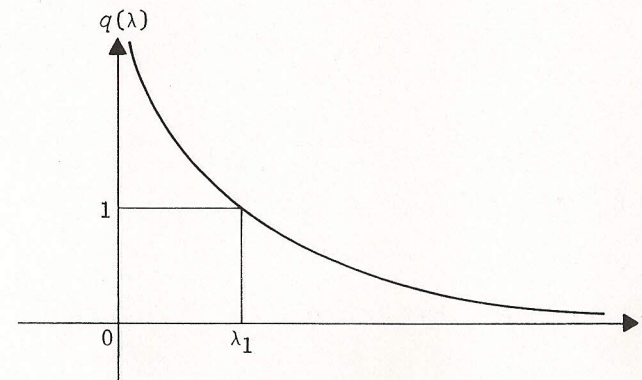


Figure 9.1

Since λ_1 is simple, its corresponding eigenspace has dimension one, and so any eigenvector corresponding to it is some multiple of x_1 . Let us summarize these results in the following theorem:

THEOREM 9.1 A Leslie matrix L has a unique positive eigenvalue λ_1 . This eigenvalue is simple and has an eigenvector x_1 all of whose entries are positive.

We shall now show that the long-term behavior of the age distribution of the population is determined by the positive eigenvalue λ_1 and its eigenvector x_1 .

In Exercise 9.9, we ask the reader to prove the following result:

THEOREM 9.2 If λ_1 is the unique positive eigenvalue of a Leslie matrix L and λ_i is any other real or complex eigenvalue of L , then $|\lambda_i| \leq \lambda_1$.

Because of Theorem 9.2, λ_1 is called a *dominant eigenvalue* of L . For our purposes we actually need more; namely, that $|\lambda_i| < \lambda_1$ for all other eigenvalues of L . In this case, we say that λ_1 is a *strictly dominant eigenvalue* of L . However, as the following example shows, not all Leslie matrices satisfy this condition.

EXAMPLE 9.2 Let

$$L = \begin{bmatrix} 0 & 0 & 6 \\ 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \end{bmatrix}.$$

Then the characteristic polynomial of L is

$$p(\lambda) = |\lambda I - L| = \lambda^3 - 1.$$

The eigenvalues of L are thus the solutions of $\lambda^3 = 1$; namely,

$$\lambda = 1, -\frac{1}{2} + \frac{\sqrt{3}}{2}i, -\frac{1}{2} - \frac{\sqrt{3}}{2}i.$$

All three eigenvalues have absolute value one, and so the unique positive eigenvalue $\lambda_1 = 1$ is not strictly dominant. Note that this matrix has the property that $L^3 = I$. This means that for any choice of the initial age distribution $x^{(0)}$, we have

$$x^{(0)} = x^{(3)} = x^{(6)} = \dots = x^{(3k)} = \dots$$

The age distribution vector thus oscillates with a period of three time units. Such oscillations (or *population waves*, as they are called) could not occur if λ_1 were strictly dominant, as we shall see below.

It is beyond the scope of this book to discuss necessary and sufficient conditions for λ_1 to be a strictly dominant eigenvalue. However, we will state the following sufficient condition without proof:

THEOREM 9.3 If two successive entries a_i and a_{i+1} in the first row of a Leslie matrix L are nonzero, then the positive eigenvalue of L is strictly dominant.

Thus, if the female population has two successive fertile age classes, then its Leslie matrix has a strictly dominant eigenvalue. This is always the case for realistic populations if the duration of the age classes is sufficiently small. Notice that in Example 9.2 there is only one fertile age class (the third), and so the condition of Theorem 9.3 is not satisfied. In what follows, we shall always assume that the condition of Theorem 9.3 is satisfied.

Let us assume that L is diagonalizable. This is not really necessary for the conclusions we shall draw, but it does simplify the arguments. In this case, L has n eigenvalues, $\lambda_1, \lambda_2, \dots, \lambda_n$, not necessarily distinct, and n linearly independent eigenvectors, x_1, x_2, \dots, x_n , corresponding to them. In this listing we place the strictly dominant eigenvalue λ_1 first. We construct a matrix P whose columns are the eigenvectors of L :

$$P = \begin{bmatrix} x_1 & | & x_2 & | & x_3 & | & \dots & | & x_n \end{bmatrix}.$$

The diagonalization of L is then given by the equation

$$L = P \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix} P^{-1}.$$

From this it follows that

$$L^k = P \begin{bmatrix} \lambda_1^k & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^k & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n^k \end{bmatrix} P^{-1},$$

for $k=1, 2, \dots$. For any initial age distribution vector $\mathbf{x}^{(0)}$ we then have

$$L^k \mathbf{x}^{(0)} = P \begin{bmatrix} \lambda_1^k & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^k & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n^k \end{bmatrix} P^{-1} \mathbf{x}^{(0)}$$

for $k=1, 2, \dots$. Dividing both sides of this equation by λ_1^k and using the fact that $\mathbf{x}^{(k)} = L^k \mathbf{x}^{(0)}$, we have

$$\frac{1}{\lambda_1^k} \mathbf{x}^{(k)} = P \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & (\lambda_2/\lambda_1)^k & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & (\lambda_n/\lambda_1)^k \end{bmatrix} P^{-1} \mathbf{x}^{(0)}. \quad (9.9)$$

Since λ_1 is the strictly dominant eigenvalue, $|\lambda_i/\lambda_1| < 1$ for $i=2, 3, \dots, n$. It follows that

$$(\lambda_i/\lambda_1)^k \rightarrow 0 \text{ as } k \rightarrow \infty \text{ for } i=2, 3, \dots, n.$$

Using this fact, we may take the limit of both sides of (9.9) to obtain

$$\lim_{k \rightarrow \infty} \left\{ \frac{1}{\lambda_1^k} \mathbf{x}^{(k)} \right\} = P \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} P^{-1} \mathbf{x}^{(0)}. \quad (9.10)$$

Let us denote the first entry of the column vector $P^{-1} \mathbf{x}^{(0)}$ by the constant c . As we ask the reader to show in Exercise 9.4, the righthand side of (9.10) can be written as $c \mathbf{x}_1$, where c is a positive constant which depends only on the initial age distribution vector $\mathbf{x}^{(0)}$. Thus (9.10) becomes

$$\lim_{k \rightarrow \infty} \left\{ \frac{1}{\lambda_1^k} \mathbf{x}^{(k)} \right\} = c \mathbf{x}_1. \quad (9.11)$$

Equation (9.11) gives us the approximation

$$\mathbf{x}^{(k)} \approx c \lambda_1^k \mathbf{x}_1 \quad (9.12)$$

for large values of k . From (9.12) we also have

$$\mathbf{x}^{(k-1)} \approx c \lambda_1^{k-1} \mathbf{x}_1. \quad (9.13)$$

Comparing Eqs. (9.12) and (9.13), we see that

$$\mathbf{x}^{(k)} \approx \lambda_1 \mathbf{x}^{(k-1)} \quad (9.14)$$

for large values of k . This means that for large values of time each age distribution vector is a scalar multiple of the preceding age distribution vector, the scalar being the positive eigenvalue of the Leslie matrix. Consequently, the *proportion* of females in each of the age classes becomes constant. As we shall see in the following example, these limiting proportions can be determined from the eigenvector \mathbf{x}_1 .

EXAMPLE 9.1 (REVISITED) The Leslie matrix in Example 9.1 was

$$L = \begin{bmatrix} 0 & 4 & 3 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{bmatrix}.$$

Its characteristic polynomial is $p(\lambda) = \lambda^3 - 2\lambda - 3/8$, and the reader can verify that the positive eigenvalue is $\lambda_1 = 3/2$. From (9.8) the corresponding eigenvector x_1 is

$$x_1 = \begin{bmatrix} 1 \\ b_1/\lambda_1 \\ b_1 b_2/\lambda_1^2 \end{bmatrix} = \begin{bmatrix} 1 \\ (1/2)/(3/2) \\ (1/2)(1/4)/(3/2)^2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1/3 \\ 1/18 \end{bmatrix}.$$

From (9.14) we have

$$x^{(k)} = (3/2)x^{(k-1)}$$

for large values of k . Hence, every five years the number of females in each of the three classes will increase by about 50%, as will the total number of females in the population.

From (9.12) we have

$$x^{(k)} = c(3/2)^k \begin{bmatrix} 1 \\ 1/3 \\ 1/18 \end{bmatrix}.$$

Consequently, eventually the females will be distributed among the three age classes in the ratios 1 : 1/3 : 1/18. This corresponds to a distribution of 72% of the females in the first age class, 24% of the females in the second age class, and 4% of the females in the third age class.

EXAMPLE 9.3 In this example we shall use birth and death parameters from the year 1965 for Canadian females. Since few women over 50 years of age bear children, we shall restrict ourselves to the portion of the female population between 0 and 50 years of age. The data are for 5-year age classes, so there are a total of ten age classes. Rather than write out the 10×10 Leslie matrix in full, we list the birth and death parameters as follows:

Age Interval	a_i	b_i
[0, 5)	0.00000	0.99651
[5, 10)	0.00024	0.99820
[10, 15)	0.05861	0.99802
[15, 20)	0.28608	0.99729
[20, 25)	0.44791	0.99694
[25, 30)	0.36399	0.99621
[30, 35)	0.22259	0.99460
[35, 40)	0.10457	0.99184
[40, 45)	0.02826	0.98700
[45, 50)	0.00240	—

Using numerical techniques, the positive eigenvalue and corresponding eigenvector turn out to be

$$\lambda_1 = 1.07622 \quad \text{and} \quad x_1 =$$

$$\begin{bmatrix} 1.00000 \\ 0.92594 \\ 0.85881 \\ 0.79641 \\ 0.73800 \\ 0.68364 \\ 0.63281 \\ 0.58482 \\ 0.53897 \\ 0.49429 \end{bmatrix}.$$

Thus, if Canadian women continued to reproduce and die as they did in 1965, eventually every five years their numbers would increase by 7.622%. From the eigenvector x_1 , we see that, in the limit, for every 100,000 females between 0 and 5 years of age, there will be 92,594 females between 5 and 10 years of age, 85,881 females between 10 and 15 years of age, and so forth.

Let us look again at Eq. (9.12) which gives the age distribution vector of the population for large times:

$$x^{(k)} = c\lambda_1^k x_1. \quad (9.15)$$

Three cases arise according to the value of the positive eigenvalue λ_1 :

- (i) The population is eventually increasing if $\lambda_1 > 1$.
- (ii) The population is eventually decreasing if $\lambda_1 < 1$.
- (iii) The population stabilizes if $\lambda_1 = 1$.

The case $\lambda_1 = 1$ is particularly interesting since it determines a population which has *zero population growth*. For any initial age distribution, the population approaches a limiting age distribution which is some multiple of the eigenvector x_1 . From Eqs. (9.6) and (9.7), we see that $\lambda_1 = 1$ is an eigenvalue if and only if

$$a_1 + a_2 b_1 + a_3 b_1 b_2 + \cdots + a_n b_1 b_2 \cdots b_{n-1} = 1. \quad (9.16)$$

The expression

$$R = a_1 + a_2 b_1 + a_3 b_1 b_2 + \cdots + a_n b_1 b_2 \cdots b_{n-1} \quad (9.17)$$

is called the *net reproduction rate* of the population. (See Exercise 9.5 for a demographic interpretation of R .) Thus we can say that a population has zero population growth if and only if its net reproduction rate is one.

EXERCISES

9.1 Suppose a certain animal population is divided into two age classes and has a Leslie matrix

$$L = \begin{bmatrix} 1 & 3/2 \\ 1/2 & 0 \end{bmatrix}.$$

- (a) Calculate the positive eigenvalue λ_1 of L and the corresponding eigenvector x_1 .
- (b) Beginning with the initial age distribution vector

$$x^{(0)} = \begin{bmatrix} 100 \\ 0 \end{bmatrix}$$

calculate $x^{(1)}$, $x^{(2)}$, $x^{(3)}$, $x^{(4)}$, and $x^{(5)}$, rounding off to the nearest integer when necessary.

- (c) Calculate $x^{(6)}$ using the exact formula $x^{(6)} = Lx^{(5)}$ and the approximate formula $x^{(6)} \approx \lambda_1 x^{(5)}$.

- 9.2 Find the characteristic polynomial of a general Leslie matrix given by Eq. (9.4).
- 9.3 Show that the positive eigenvalue λ_1 of a Leslie matrix is always simple. Recall that a root λ_0 of a polynomial $q(\lambda)$ is simple if and only if $q'(\lambda_0) \neq 0$.
- 9.4 Show that the righthand side of Eq. (9.10) is cx_1 where c is the first entry of the column vector $P^{-1}x^{(0)}$.
- 9.5 Show that the net reproduction rate R , defined by (9.17), can be interpreted as the average number of daughters born to a single female during her expected lifetime.
- 9.6 Show that a population is eventually decreasing if and only if its net reproduction rate is less than one. Similarly, show that a population is eventually increasing if and only if its net reproduction rate is greater than one.
- 9.7 Calculate the net reproduction rate of the animal population in Example 9.1.
- 9.8 (For readers with a hand calculator) Calculate the net reproduction rate of the Canadian female population in Example 9.3.
- 9.9 (For readers who have had a course in Complex Variables) Prove Theorem 9.2. Hint: write $\lambda_i = re^{i\theta}$, substitute into (9.7), take the real parts of both sides, and show that $r \leq \lambda_1$.

Harvesting of Animal Populations

The Leslie matrix model of population growth is used to model the sustainable harvesting of an animal population. The effect of harvesting different fractions of different age groups is investigated.

PREREQUISITES: Chapter 9: Age-Specific Population Growth

INTRODUCTION

In Chapter 9, the Leslie matrix model for the growth of a female population divided into discrete age classes was described. In this chapter, we shall investigate the effects of harvesting an animal population growing according to such a model. By *harvesting* we mean the removal of animals from the population. The word "harvesting" is not necessarily a euphemism for "slaughtering"; the animals may be removed from the population for other purposes.

We shall restrict ourselves to what are called *sustainable harvesting policies*. By this we mean the following:

Sustainable Harvesting Policy

A harvesting policy in which an animal population is periodically harvested is said to be sustainable if the yield of each harvest is the same and the age distribution of the population remaining after each harvest is the same.

Thus, the animal population is not depleted by a sustainable harvesting policy; only the excess growth is exploited.

As in Chapter 9, we shall only discuss the females of the population. If the number of males in each age class is equal to the number of females — a reasonable assumption for many populations — then our harvesting policies will also apply to the male portion of the population.

THE HARVESTING MODEL

Figure 10.1 illustrates the basic idea of the model. We begin with a population having a particular age distribution. It undergoes a growth period which will be described by the Leslie matrix. At the end of the growth period, a certain fraction of each age class is harvested. The duration of the harvest is to be short in comparison with the growth period so that any growth or change in the population during the harvest period may be neglected. Finally, the population left unharvested is to have the same age distribution as the original population. This cycle repeats after each harvest, so that the yield is sustainable.

To describe this harvesting model mathematically, let

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

be the age distribution vector of the population at the beginning of the growth period. Thus x_i is the number of females in the i -th class left unharvested. As in Chapter 9, we require that the duration of each age class be identical with the duration of the growth period. For example, if the population is harvested once a year, then the population is to be divided into one-year age classes.

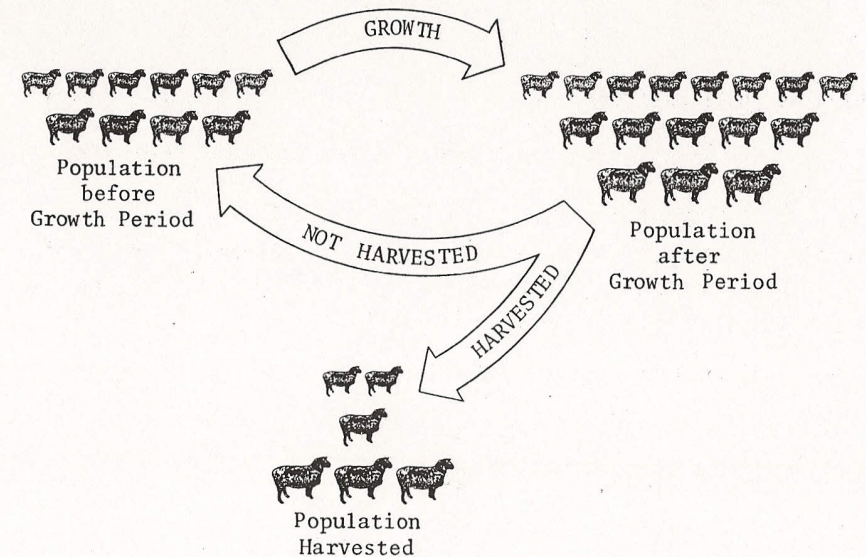


Figure 10.1

If L is the Leslie matrix describing the growth of the population, then the vector Lx is the age distribution vector of the population at the end of the growth period, immediately before the periodic harvest. Let h_i , for $i = 1, 2, \dots, n$, be the fraction of females from the i -th class which is harvested. We use these n numbers to form an $n \times n$ diagonal matrix

$$H = \begin{bmatrix} h_1 & 0 & 0 & \cdots & 0 \\ 0 & h_2 & 0 & \cdots & 0 \\ 0 & 0 & h_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & h_n \end{bmatrix}$$

which we shall call the *harvesting matrix*. By definition, we have

$$0 \leq h_i \leq 1 \quad \text{for } i = 1, 2, \dots, n.$$

That is, we may harvest none ($h_i = 0$), all ($h_i = 1$), or some proper fraction ($0 < h_i < 1$) of each of the n classes. Since the number of females in the i -th class immediately before each harvest is the i -th entry $(Lx)_i$ of the vector Lx , it can be seen that the i -th entry of the column vector

$$HLx = \begin{bmatrix} h_1(Lx)_1 \\ h_2(Lx)_2 \\ \vdots \\ h_n(Lx)_n \end{bmatrix}$$

is the number of females harvested from the i -th class.

From the definition of a sustainable harvesting policy, we have

$$\begin{pmatrix} \text{age distribution} \\ \text{at end of} \\ \text{growth period} \end{pmatrix} - \begin{pmatrix} \text{harvest} \end{pmatrix} = \begin{pmatrix} \text{age distribution} \\ \text{at beginning of} \\ \text{growth period} \end{pmatrix}$$

or mathematically,

$$Lx - HLx = x. \quad (10.1)$$

If we write Eq. (10.1) in the form

$$(I - H)Lx = x \quad (10.2)$$

we see that x must be an eigenvector of the matrix $(I - H)L$ corresponding to the eigenvalue one. As we shall now describe, this places certain restrictions on the values of h_i and x .

Suppose the Leslie matrix of the population is

$$L = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_{n-1} & a_n \\ b_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & b_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & b_{n-1} & 0 \end{bmatrix}. \quad (10.3)$$

Then the matrix $(I - H)L$ is easily computed:

$$(I - H)L = \begin{bmatrix} (1 - h_1)a_1 & (1 - h_1)a_2 & (1 - h_1)a_3 & \cdots & (1 - h_1)a_{n-1} & (1 - h_1)a_n \\ (1 - h_2)b_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & (1 - h_3)b_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & (1 - h_n)b_{n-1} & 0 \end{bmatrix}.$$

We thus see that $(I - H)L$ is a matrix which has the same mathematical form as a Leslie matrix. In Chapter 9, we showed that a necessary and sufficient condition for a Leslie matrix to have one as an eigenvalue is that its net reproduction rate also be one. (See Eq. (9.16) on page 122.) Calculating the net reproduction rate of $(I - H)L$ and setting it equal to one, we obtain (verify):

$$(1 - h_1)[a_1 + a_2 b_1 (1 - h_2) + a_3 b_1 b_2 (1 - h_2)(1 - h_3) + \cdots + a_n b_1 b_2 \cdots b_{n-1} (1 - h_2)(1 - h_3) \cdots (1 - h_n)] = 1. \quad (10.4)$$

This equation places a restriction on the allowable harvesting fractions. Only those values of h_1, h_2, \dots, h_n which satisfy Eq. (10.4) and which lie in the interval $[0, 1]$ can produce a sustainable yield.

If h_1, h_2, \dots, h_n do satisfy (10.4), then the matrix $(I - H)L$ has the desired eigenvalue $\lambda_1 = 1$; and furthermore this eigenvalue has multiplicity one since the positive eigenvalue of a Leslie matrix always has multiplicity one (Theorem 9.1, page 116). This means that there is only one linearly independent eigenvector x satisfying Eq. (10.2). As in Chapter 9 (Eq. 9.8, page 115), we pick the following normalized eigenvector:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ b_1(1-h_2) \\ b_1b_2(1-h_2)(1-h_3) \\ b_1b_2b_3(1-h_2)(1-h_3)(1-h_4) \\ \vdots \\ b_1b_2b_3 \cdots b_{n-1}(1-h_2)(1-h_3) \cdots (1-h_n) \end{bmatrix}. \quad (10.5)$$

Any other solution \mathbf{x} of (10.2) is a multiple of \mathbf{x}_1 . The vector \mathbf{x}_1 thus determines the proportion of females within each of the n classes after a harvest under a sustainable harvesting policy. But there is an ambiguity in the total number of females in the population after each harvest. This can be determined by some auxiliary condition, such as an ecological or economic constraint. For example, for a population economically supported by the harvester, the largest population the harvester can afford to raise between harvests would determine the particular constant \mathbf{x}_1 is multiplied by to produce the appropriate vector \mathbf{x} in (10.2). For a wild population — deer, whales, bears, etc. — the natural habitat of the population would determine how large the total population may be between harvests.

Summarizing our results so far, we see that there is a wide choice in the values of h_1, h_2, \dots, h_n which will produce a sustainable yield. But once these values are selected, the proportional age distribution of the population after each harvest is uniquely determined by the normalized eigenvector \mathbf{x}_1 defined by Eq. (10.5). Let us now consider a few particular harvesting strategies of this type.

UNIFORM HARVESTING

With many populations it is difficult to distinguish or catch animals of specific ages. If animals are caught at random, we may reasonably assume the same fraction of each age class is harvested. Let us therefore set

$$h = h_1 = h_2 = \cdots = h_n.$$

Equation (10.2) then reduces to (verify):

$$L\mathbf{x} = \left(\frac{1}{1-h} \right) \mathbf{x}.$$

Hence, $1/(1-h)$ must be the unique positive eigenvalue λ_1 of the Leslie growth matrix L . That is,

$$\lambda_1 = \frac{1}{1-h}.$$

Solving for the harvesting fraction h , we obtain

$$h = 1 - 1/\lambda_1. \quad (10.6)$$

The vector \mathbf{x}_1 , in this case, is the same as the eigenvector of L corresponding to the eigenvalue λ_1 . From Chapter 9 (Eq. (9.8), page 115), this is

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ b_1/\lambda_1 \\ b_1b_2/\lambda_1^2 \\ b_1b_2b_3/\lambda_1^3 \\ \vdots \\ b_1b_2 \cdots b_{n-1}/\lambda_1^{n-1} \end{bmatrix}. \quad (10.7)$$

From Eq. (10.6), we can see that the larger λ_1 is, the larger is the fraction of animals we can harvest without depleting the population. We also notice that we need $\lambda_1 > 1$ in order that the harvesting fraction h lie in the interval $(0, 1]$. This is to be expected since $\lambda_1 > 1$ is the condition that the population be increasing.

EXAMPLE 10.1 For a certain species of domestic sheep in New Zealand with a growth period of one year, the following Leslie matrix was found (G. Caughley, "Parameters for Seasonally Breeding Populations," *Ecology*, Vol. 48, 1967, pages 834-839):

$$L = \begin{bmatrix} .000 & .045 & .391 & .472 & .484 & .546 & .543 & .502 & .468 & .459 & .433 & .421 \\ .845 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & .975 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .965 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .950 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & .926 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & .895 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .850 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .786 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .691 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .561 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .370 & 0 \end{bmatrix}$$

The sheep have a lifespan of 12 years so that they are divided into 12 age classes of duration one year each. By the use of numerical techniques, the unique positive eigenvalue of L can be found to be

$$\lambda_1 = 1.221 .$$

From Eq. (10.6), the harvesting fraction h is

$$h = 1 - 1/\lambda_1 = 1 - 1/1.221 = 0.181 .$$

Thus, the uniform harvesting policy is one in which 18.1% of the sheep from each of the 12 age classes is harvested every year. From Eq. (10.7), the age distribution vector of the sheep after each harvest is proportional to

$$x_1 = \begin{bmatrix} 1.000 \\ 0.692 \\ 0.552 \\ 0.436 \\ 0.339 \\ 0.257 \\ 0.189 \\ 0.131 \\ 0.084 \\ 0.048 \\ 0.022 \\ 0.007 \end{bmatrix} . \tag{10.8}$$

From (10.8) we see that for every 1000 sheep between 0 and 1 years of age which are not harvested, there are 692 sheep between 1 and 2 years of age, 552 sheep between 2 and 3 years of age, and so forth.

HARVESTING ONLY THE YOUNGEST AGE CLASS

In some populations, only the youngest females are of any economic value, and so the harvester seeks to harvest only the females from the youngest age class. Accordingly, let us set

$$\begin{aligned} h_1 &= h, \\ h_2 = h_3 = \dots = h_n &= 0. \end{aligned}$$

Equation (10.4) then reduces to

$$(1 - h)(a_1 + a_2 b_1 + a_3 b_1 b_2 + \dots + a_n b_1 b_2 \dots b_{n-1}) = 1,$$

or

$$(1 - h)R = 1$$

where R is the net reproduction rate of the population. (See Eq. (9.17), page 122.) Solving for h , we obtain

$$h = 1 - 1/R. \tag{10.9}$$

We notice from this equation that only if $R > 1$ is a sustainable harvesting policy possible. This is reasonable since only if $R > 1$ is the population increasing. From Eq. (10.5), the age distribution vector after each harvest is proportional to the vector

$$x_1 = \begin{bmatrix} 1 \\ b_1 \\ b_1 b_2 \\ b_1 b_2 b_3 \\ \vdots \\ b_1 b_2 b_3 \dots b_{n-1} \end{bmatrix} . \tag{10.10}$$

EXAMPLE 10.2 Let us apply this type of sustainable harvesting policy to the sheep population in Example 10.1. For the net reproduction rate of the population, we find:

$$\begin{aligned}
 R &= a_1 + a_2 b_1 + a_3 b_1 b_2 + \dots + a_n b_1 b_2 \dots b_{n-1} \\
 &= (.000) + (.045)(.845) + \dots + (.421)(.845)(.975) \dots (.370) \\
 &= 2.513.
 \end{aligned}$$

From Eq. (10.9), the fraction of the first age class harvested is

$$h = 1 - 1/R = 1 - 1/2.513 = .602 .$$

From Eq. (10.10), the age distribution of the sheep population after the harvest is proportional to the vector

$$\mathbf{x}_1 = \begin{bmatrix} 1.000 \\ 0.845 \\ (.845)(.975) \\ (.845)(.975)(.965) \\ \vdots \\ (.845)(.975) \dots (.370) \end{bmatrix} = \begin{bmatrix} 1.000 \\ 0.845 \\ 0.824 \\ 0.795 \\ 0.755 \\ 0.699 \\ 0.626 \\ 0.532 \\ 0.418 \\ 0.289 \\ 0.162 \\ 0.060 \end{bmatrix} \quad (10.11)$$

A direct calculation gives us the following (see also Exercise (10.3):

$$L\mathbf{x}_1 = \begin{bmatrix} 2.513 \\ 0.845 \\ 0.824 \\ 0.795 \\ 0.755 \\ 0.699 \\ 0.626 \\ 0.532 \\ 0.418 \\ 0.289 \\ 0.162 \\ 0.060 \end{bmatrix} \quad (10.12)$$

The vector $L\mathbf{x}_1$ is the age distribution vector immediately before the harvest. The total of all entries in $L\mathbf{x}_1$ is 8.518, so that the first entry 2.513 is 29.5% of the total. This means that immediately before each harvest, 29.5% of the population is in the youngest age class. Since 60.2% of this class is harvested, it follows that 17.8% (= 60.2% of 29.5%) of the entire sheep population is harvested each year. This can be compared with the uniform harvesting policy of Example 10.1, in which 18.1% of the sheep population is harvested each year.

OPTIMAL SUSTAINABLE YIELD

We saw in Example 10.1 that a sustainable harvesting policy in which the same fraction of each age class is harvested produces a yield of 18.1% of the sheep population. In Example 10.2, we saw that if only the youngest age class is harvested, the resulting yield is 17.8% of the population. There are many other possible sustainable harvesting policies, and each will provide a generally different yield. It would be of interest to find a sustainable harvesting policy which produces the largest possible yield. Such a policy is called an *optimal sustainable harvesting policy* and the resulting yield is called the *optimal sustainable yield*. However, the determination of the optimal sustainable yield requires Linear Programming theory, and we cannot discuss it in detail in this chapter. (See Chapter 13 for an introduction to Linear Programming theory.) We shall, though, state the following result for reference (C. Rorres, "Optimal Sustainable Yield of a Renewable Resource," *Biometrics*, Vol. 32, 1976, pages 945 - 948):

Optimal Sustainable Yield

An optimal sustainable harvesting policy is one in which either one or two age classes are harvested. If two age classes are harvested, then the older age class is completely harvested.

As an illustration, it can be shown using Linear Programming that the optimal sustainable yield of the sheep population is attained when

$$\begin{aligned}
 h_1 &= 0.522 \\
 h_9 &= 1.000
 \end{aligned} \quad (10.13)$$

and all other values of h_i are zero. Thus, 52.2% of the sheep between 0 and 1 years of age and all of the sheep between 8 and 9 years of age are harvested. As we ask the reader to show in Exercise 10.2, the resulting optimal sustainable yield is 19.9% of the population.

EXERCISES

10.1 Let a certain animal population be divided into three one-year age classes and have as its Leslie matrix

$$L = \begin{bmatrix} 0 & 4 & 3 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{bmatrix}$$

- (a) Find the yield and the age distribution vector after each harvest if the same fraction of each of the three age classes is harvested every year.
- (b) Find the yield and the age distribution vector after each harvest if only the youngest age class is harvested every year. Also, find the fraction of the youngest age class which is harvested.

10.2 For the optimal sustainable harvesting policy described by Eqs. (10.13), find the vector x_1 which specifies the age distribution of the population after each harvest. Also, calculate the vector Lx_1 and verify that the optimal sustainable yield is 19.9% of the population.

10.3 If only the first age class of an animal population is harvested, use Eq. (10.10) to show that

$$Lx_1 - x_1 = \begin{bmatrix} R - 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where R is the net reproduction rate of the population.

10.4 If only the I -th class of an animal population is to be periodically harvested ($I = 1, 2, \dots, n$), find the corresponding harvesting fraction h_I .

10.5 Suppose all of the J -th class and a certain fraction h_I of the I -th class of an animal population is to be periodically harvested ($1 \leq I < J \leq n$). Calculate h_I .

Least Squares Fitting to Data

A technique of best fitting a line or other polynomial curve to a set of experimentally determined points in the plane is described.

PREREQUISITES: Linear systems
Matrices
Euclidean space R^n
Column space of a matrix

INTRODUCTION

A common problem in experimental work is to obtain a mathematical relationship $y = f(x)$ between two variables x and y by "fitting" a curve to points in the plane corresponding to various experimentally determined values of x and y , say

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

Based on theoretical considerations, or simply on the pattern of the points, one decides on the general form of the curve $y = f(x)$ to be fitted. Some possibilities are (Fig. 11.1):

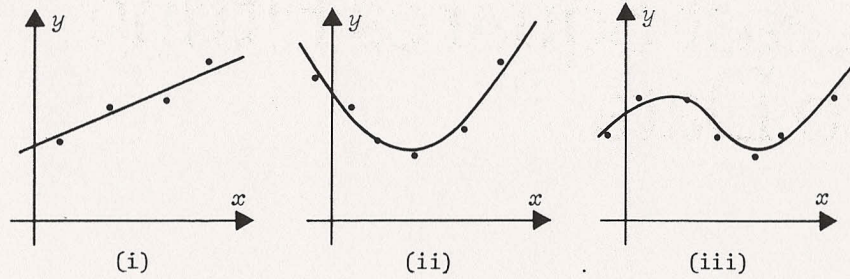


Figure 11.1

- (i) A straight line — $y = a + bx$
- (ii) A quadratic polynomial — $y = a + bx + cx^2$
- (iii) A cubic polynomial — $y = a + bx + cx^2 + dx^3$.

Because the points are obtained experimentally, there is usually some "error" in the data making it impossible to find a curve of the desired form that passes through all of the points. Thus, the idea is to choose the curve (by determining its coefficients) which "best" fits the data. We begin with the simplest case: fitting a straight line to the data points.

LEAST SQUARES FITTING OF A STRAIGHT LINE

Suppose we want to fit a straight line

$$y = a + bx$$

to the experimentally determined points

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

If the data points are collinear, the line would pass through all n points, and so the unknown coefficients a and b would satisfy

$$\begin{aligned} y_1 &= a + bx_1 \\ y_2 &= a + bx_2 \\ &\vdots \\ y_n &= a + bx_n. \end{aligned}$$

We may write this system in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

or more compactly as

$$y = Mv \tag{11.1}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad M = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad v = \begin{bmatrix} a \\ b \end{bmatrix}. \tag{11.2}$$

If the data points are not collinear, it is impossible to find coefficients a and b that satisfy (11.1) exactly; thus no matter how we choose v , the difference

$$y - Mv$$

between the two sides of (11.1) will not be zero. In this case, our objective will be to find a vector v which minimizes the Euclidean length of this difference

$$\|y - Mv\|.$$

If $v^* = \begin{bmatrix} a^* \\ b^* \end{bmatrix}$ is such a minimizing vector, we call the line $y = a^* + b^*x$ the *least squares straight line fit* to the data.

To see how to find such a vector v^* , refer to Figure 11.2. The vector y is a fixed vector in R^n . And as v varies over all possible

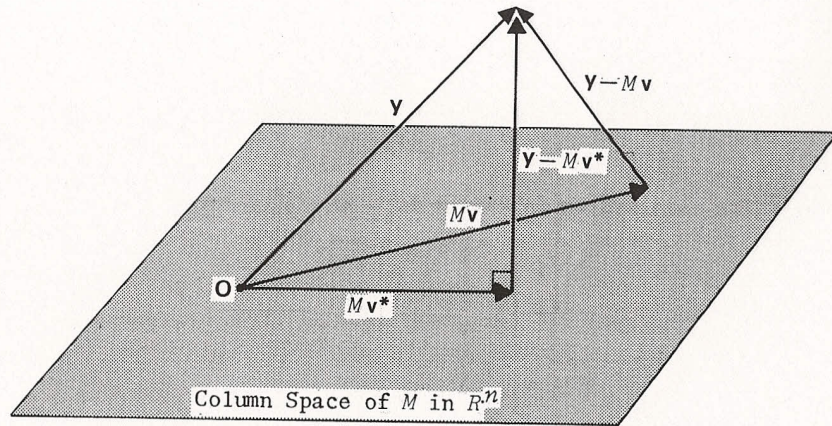


Figure 11.2

values, the vectors Mv form a subspace of R^n — the range space or column space of the matrix M . As suggested by Fig. 11.2, if $y - Mv$ is to have minimum length it must be orthogonal to the range space. Let v^* be that vector such that $y - Mv^*$ is orthogonal to the range space. This means that the Euclidean inner product of $y - Mv^*$ with Mv is zero for any vector v . Thus,

$$(Mv)^t (y - Mv^*) = 0$$

for all vectors v . Or, since $(Mv)^t = v^t M^t$,

$$v^t (M^t y - M^t M v^*) = 0$$

for all vectors v . But this equation states that the fixed vector $M^t y - M^t M v^*$ is orthogonal to every vector v . This is possible only if this fixed vector is the zero vector; that is, if

$$M^t M v^* = M^t y. \tag{11.3}$$

Since M is an $n \times 2$ matrix (see Eq. (11.2)), the matrix $M^t M$ is a 2×2 matrix. Thus if $M^t M$ is invertible, (11.3) has the unique solution

$$v^* = (M^t M)^{-1} M^t y.$$

REMARK In Exercise 11.6 we ask the reader to show that $M^t M$ fails to be invertible if and only if the n data points lie on a vertical line in the xy -plane.

Using this remark, we may summarize our results as follows (where we now write v instead of v^* to simplify the notation):

Given $n > 1$ points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, not all lying on a vertical line, the least squares straight line fit to these points,

$$y = a + bx,$$

has coefficients given by

$$v = (M^t M)^{-1} M^t y$$

where

$$v = \begin{bmatrix} a \\ b \end{bmatrix}, \quad M = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \text{and} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

EXAMPLE 11.1 Find the least squares straight line fit to the four points $(0,1), (1,3), (2,4)$, and $(3,4)$.

SOLUTION We have

$$M = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix},$$

$$M^t M = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix},$$

$$(M^t M)^{-1} = \frac{1}{10} \begin{bmatrix} 7 & -3 \\ -3 & 2 \end{bmatrix},$$

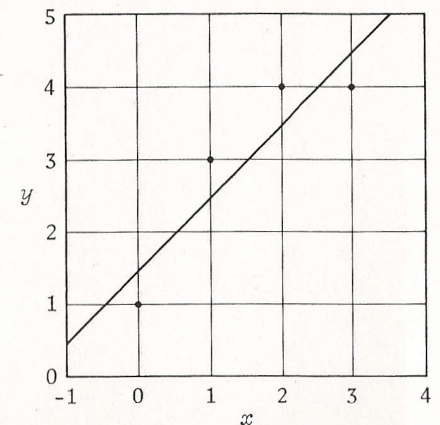


Figure 11.3

$$v = (M^t M)^{-1} M^t y = \frac{1}{10} \begin{bmatrix} 7 & -3 \\ -3 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}$$

And so the desired line is $y = 1.5 + x$.

Figure 11.4 provides an alternate interpretation of least squares fitting of a straight line. We note that $\|y - Mv\|$ and

$$\|y - Mv\|^2 = (y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + \dots + (y_n - a - bx_n)^2$$

are minimized by the same vector v . From Fig. 11.4 we see that $|y_i - a - bx_i|$ is the vertical distance d_i from the data point (x_i, y_i) to the line $y = a + bx$. If we think of d_i as the "vertical error" at the point (x_i, y_i) , then the least squares straight line fit minimizes

$$d_1^2 + d_2^2 + \dots + d_n^2,$$

which is the sum of the squares of the vertical errors.

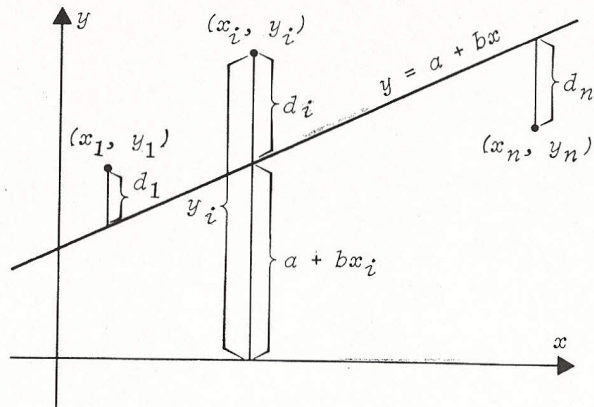


Figure 11.4

EXAMPLE 11.2 Hooke's law in Physics states that the length x of a uniform spring is a linear function of the applied force y . If we write $y = a + bx$, then the coefficient b is called the spring constant. Suppose a particular unstretched spring has a measured length of 6.1 inches (i.e. $x = 6.1$ when $y = 0$). Forces of 2 pounds, 4 pounds, and 6 pounds are then applied to the spring, and the corresponding lengths are found to be 7.6 inches, 8.7 inches, and 10.4 inches. Find the spring constant of this spring.

SOLUTION We have

$$M = \begin{bmatrix} 1 & 6.1 \\ 1 & 7.6 \\ 1 & 8.7 \\ 1 & 10.4 \end{bmatrix}, \quad y = \begin{bmatrix} 0 \\ 2 \\ 4 \\ 6 \end{bmatrix}$$

and

$$v = \begin{bmatrix} a \\ b \end{bmatrix} = (M^t M)^{-1} M^t y = \begin{bmatrix} -8.6 \\ 1.4 \end{bmatrix}$$

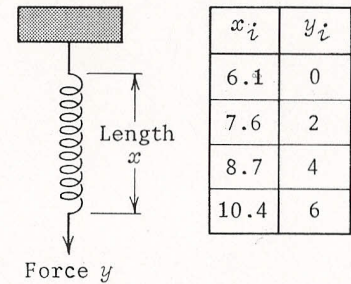


Figure 11.5

The spring constant is thus $b = 1.4$ pounds/inch.

LEAST SQUARES FITTING OF A POLYNOMIAL

The technique described for fitting a straight line to data points generalizes easily to fitting a polynomial of any specified degree to data points. Let us attempt to fit a polynomial of fixed degree m

$$y = a_0 + a_1 x + \dots + a_m x^m \tag{11.4}$$

to n points

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

Substituting these n values of x and y into (11.4) yields the n equations

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}$$

or, more simply,

$$y = Mv, \quad (11.5)$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad M = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix}, \quad v = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}.$$

Exactly as before, it can be shown that the least squares solution for v is given by

$$v = (M^t M)^{-1} M^t y, \quad (11.6)$$

which determines the coefficients of the polynomial. (See Exercise 11.8 for a discussion of the invertibility of the matrix $M^t M$.)

EXAMPLE 11.3 According to Newton's second law of motion, a body near the earth's surface falls vertically downward according to the equation

$$s = s_0 + v_0 t + \frac{1}{2} g t^2 \quad (11.7)$$

where

s = vertical displacement downward relative to some fixed point,

s_0 = initial displacement at time $t=0$,

v_0 = initial velocity at time $t=0$,

g = acceleration of gravity at earth's surface.

Suppose a laboratory experiment is performed to evaluate g using this equation. A weight is released with unknown initial displacement and velocity and at certain times the distances fallen from some fixed reference point are measured. In particular, suppose that at times $t = .1, .2, .3, .4$, and $.5$ seconds it is found that the weight has fallen $s = -0.18, 0.31, 1.03, 2.48$, and 3.73 feet, respectively, from the reference point. Find an approximate value of g using these data.

SOLUTION The mathematical problem is to fit a quadratic curve

$$s = \alpha_0 + \alpha_1 t + \alpha_2 t^2 \quad (11.8)$$

to the five data points:

$$(.1, -0.18), (.2, 0.31), (.3, 1.03), (.4, 2.48), (.5, 3.73).$$

The necessary calculations are

$$M = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ 1 & t_3 & t_3^2 \\ 1 & t_4 & t_4^2 \\ 1 & t_5 & t_5^2 \end{bmatrix} = \begin{bmatrix} 1 & .1 & .01 \\ 1 & .2 & .04 \\ 1 & .3 & .09 \\ 1 & .4 & .16 \\ 1 & .5 & .25 \end{bmatrix},$$

$$y = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{bmatrix} = \begin{bmatrix} -0.18 \\ 0.31 \\ 1.03 \\ 2.48 \\ 3.73 \end{bmatrix},$$

and

$$v = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = (M^t M)^{-1} M^t y = \begin{bmatrix} -0.40 \\ 0.35 \\ 16.1 \end{bmatrix}.$$

From (11.7) and (11.8) we have $\alpha_2 = \frac{1}{2}g$, and so

$$g = 2\alpha_2 = 2(16.1) = 32.2 \text{ feet/second}^2.$$

If desired, we may also evaluate the initial displacement and initial velocity of the weight:

$$s_0 = \alpha_0 = -0.40 \text{ feet,}$$

$$v_0 = \alpha_1 = 0.35 \text{ feet/second.}$$

In Fig. 11.6 we have plotted the five data points and the approximating polynomial.

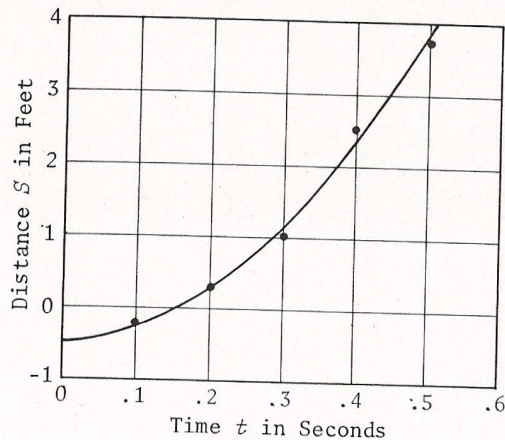


Figure 11.6

EXERCISES

- 1 11.1 Find the least squares straight line fit to the three points (0,0), (1,2), and (2,7).
- 11.2 Find the least squares straight line fit to the four points (0,1), (2,0), (3,1), and (3,2).
- 1 11.3 Find the quadratic polynomial which best fits the four points (2,0), (3,-10), (5,-48), and (6,-76).

11.4 Find the cubic polynomial which best fits the five points (-1,-14), (0,-5), (1,-4), (2,1), and (3,22).

11.5 Show that if M is an $m \times n$ matrix with linearly independent columns, then $M^t M$ is an $n \times n$ invertible matrix.

11.6 Show that the matrix M in (11.2) has linearly independent columns if and only if at least two of the numbers x_1, x_2, \dots, x_n are distinct. Conclude from this and from Exercise 11.5 that the matrix $M^t M$ in (11.3) is invertible if and only if the n data points do not lie on a vertical line in the xy -plane.

11.7 Show that the columns of the $n \times (m+1)$ matrix M in (11.5) are linearly independent if $n > m$ and at least $m+1$ of the numbers x_1, x_2, \dots, x_n are distinct.

11.8 Using Exercise 11.7, show that a sufficient condition that the matrix $M^t M$ in (11.6) is invertible is that $n > m$ and at least $m+1$ of the numbers x_1, x_2, \dots, x_n are distinct.

11.9 The owner of a rapidly expanding business finds that for the first five months of the year his sales are \$4.0, 4.4, 5.2, 6.4, and 8.0 thousand. He plots these figures on a graph and conjectures that for the rest of the year his sales curve can be approximated by a quadratic polynomial. Find the least squares quadratic polynomial fit to the sales curve and use it to project the sales for the 12th month of the year.

A Least Squares Model for Human Hearing: Fourier Series

The method of least squares approximation is motivated by energy considerations in a model for human hearing. Linear Algebra techniques are used to approximate a function over an interval by a linear combination of sine and cosine terms.

PREREQUISITES: Inner product spaces
Orthogonal projection
 $C[a, b]$ (the vector space of continuous functions on $[a, b]$)

INTRODUCTION

We begin with a brief discussion of the nature of sound and human hearing. Figure 12.1 is a schematic diagram of the ear showing its three main components: the outer ear, middle ear, and inner ear. Sound waves enter the outer ear where they are channeled to the eardrum, causing it to vibrate. Three tiny bones in the middle ear mechanically link the eardrum with the snail-shaped cochlea within the inner ear. These bones pass on the vibrations of the eardrum to a fluid within the cochlea. The cochlea contains thou-

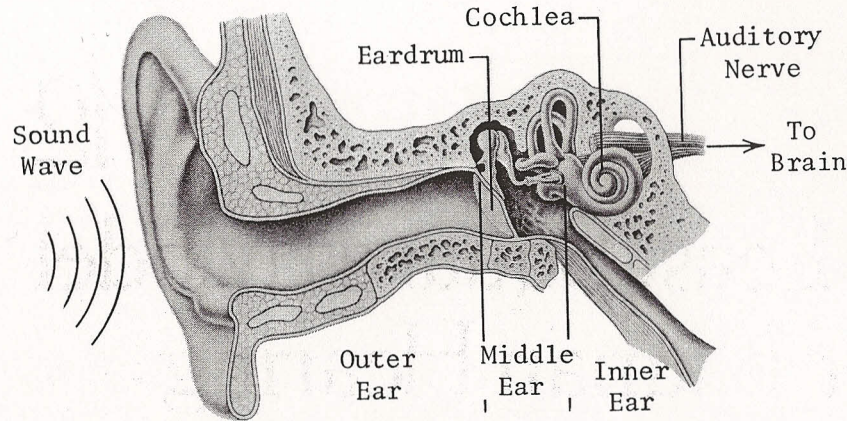


Figure 12.1

sands of minute hairs which oscillate with the fluid. Those near the entrance of the cochlea are stimulated by high frequencies and those near the tip are stimulated by low frequencies. The movements of these hairs activate nerve cells which send signals along various neural pathways to the brain, where the signals are interpreted as sound.

The sound waves themselves are variations in time of the air pressure. For the auditory system, the most elementary type of sound wave is a sinusoidal variation in the air pressure. This type of sound wave stimulates the hairs within the cochlea in such a way that a nerve impulse along a single neural pathway is produced (Fig. 12.2). A sinusoidal sound wave may be described by a function of time

$$q(t) = A_0 + A \sin \omega(t - \delta) \quad (12.1)$$

where $q(t)$ is the atmospheric pressure at the eardrum, A_0 is the normal atmospheric pressure, A is the maximum deviation of the pressure from the normal atmospheric pressure, $\omega/2\pi$ is the frequency of the wave in cycles per second, and δ is the phase angle of the wave. To be perceived as sound, such sinusoidal waves must have frequencies within a certain range. For humans this range is roughly 20 cps to 20,000 cps. Frequencies outside of this range will not stimulate the hairs within the cochlea enough to produce nerve signals.

To a reasonable degree of accuracy, the ear is a linear system. This means that if a complex sound wave is a finite sum of sinusoidal components of different amplitudes, frequencies, and phase angles, say

$$q(t) = A_0 + A_1 \sin \omega_1(t - \delta_1) + A_2 \sin \omega_2(t - \delta_2) + \dots + A_n \sin \omega_n(t - \delta_n), \quad (12.2)$$

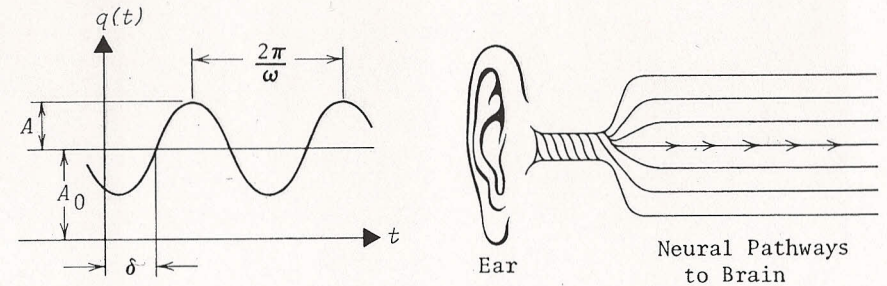


Figure 12.2

then the response of the ear consists of nerve impulses along the same neural pathways that would be stimulated by the individual components (Fig. 12.3).

Let us now consider some periodic sound wave $p(t)$ with period T (i.e., $p(t) \equiv p(t+T)$) which is *not* a finite sum of sinusoidal waves. If we examine the response of the ear to such a periodic wave, we find that it is the same as the response to some wave which is the sum of sinusoidal waves. That is, there is some sound wave $q(t)$ as given by Eq. (12.2) which produces the same response as $p(t)$, even though $p(t)$ and $q(t)$ are different functions of time.

We now want to determine the frequencies, amplitudes, and phase angles of the sinusoidal components of $q(t)$. Since $q(t)$ produces the same response as the periodic wave $p(t)$, it is reasonable to expect that $q(t)$ has the same period T as $p(t)$. This requires that each sinusoidal term in $q(t)$ have period T . Consequently, the frequencies of the sinusoidal components must be integer multiples of

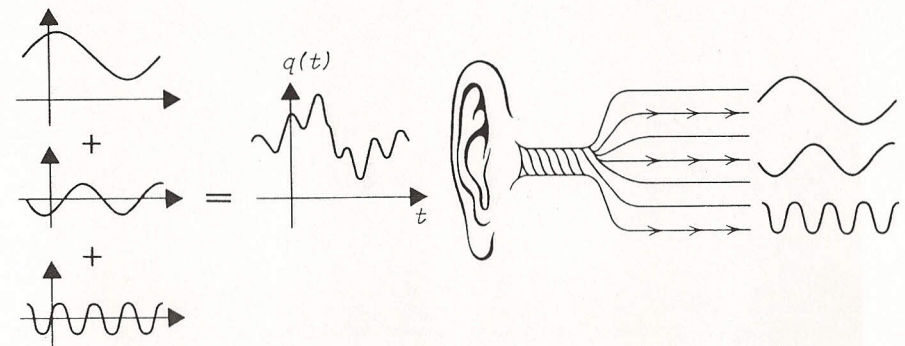


Figure 12.3

the basic frequency $1/T$ of the $p(t)$. That is, the ω_k in Eq. (12.2) must be of the form

$$\omega_k = 2k\pi/T, \quad k = 1, 2, \dots$$

But since the ear cannot perceive sinusoidal waves with frequencies greater than 20,000 cps, we may omit those values of k for which $\omega_k/2\pi = k/T$ is greater than 20,000. Thus, $q(t)$ is of the form

$$q(t) = A_0 + A_1 \sin \frac{2\pi}{T}(t - \delta_1) + \dots + A_n \sin \frac{2n\pi}{T}(t - \delta_n) \quad (12.3)$$

where n is the largest integer such that n/T is not greater than 20,000.

We now turn our attention to the values of the amplitudes A_0, A_1, \dots, A_n and the phase angles $\delta_1, \delta_2, \dots, \delta_n$ which appear in Eq. (12.3). There is some criterion by which the auditory system "picks" these values so that $q(t)$ produces the same response as $p(t)$. To examine this criterion, let us set

$$e(t) = p(t) - q(t).$$

If we consider $q(t)$ as an approximation to $p(t)$, then $e(t)$ is the error in this approximation; an error which the ear cannot perceive. In terms of $e(t)$, the criterion for the determination of the amplitudes and the phase angles is that the quantity

$$\int_0^T [e(t)]^2 dt$$

be as small as possible. We cannot go into the physiological reasons for this, but we can remark that this expression is proportional to the *acoustic energy* of the error wave $e(t)$ over one period. In other words, it is the energy of the difference between the two sound waves $p(t)$ and $q(t)$ which determine whether the ear perceives any difference between them. If this energy is as small as possible, then the two waves produce the same sensation of sound.

In general, suppose we want to approximate a function $p(t)$ by another function $q(t)$ from a certain class. If the criterion of approximation is the smallness of the integral

$$\int_0^T [p(t) - q(t)]^2 dt, \quad (12.4)$$

then we call $q(t)$ the *least squares approximation* to $p(t)$ over the interval $[0, T]$. The integral in (12.4) is called the *least squares error*, or the *mean squares error*, of the approximation. This criterion arises quite naturally in a wide variety of engineering and scientific approximation problems. Besides the acoustics problem just discussed, some other examples are

1. Let $T(x)$ be the temperature distribution in a uniform rod lying along the x -axis from $x = 0$ to $x = \ell$ (Fig. 12.4). The thermal energy in the rod is proportional to the integral

$$\int_0^{\ell} [T(x)]^2 dx.$$

The closeness of an approximation $q(x)$ to $T(x)$ can be judged according to the thermal energy of the difference of the two temperature distributions. That energy is proportional to

$$\int_0^{\ell} [T(x) - q(x)]^2 dx,$$

which is a least squares criterion.

2. Let $E(t)$ be a periodic voltage across a resistor in an electrical circuit. The electrical energy transferred to the resistor during one period T is proportional to

$$\int_0^T [E(t)]^2 dt.$$

If $q(t)$ has the same period as $E(t)$ and is to be an approximation to $E(t)$, then the criterion

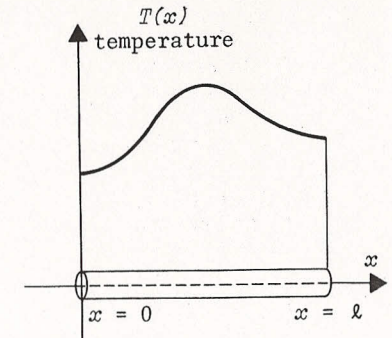


Figure 12.4

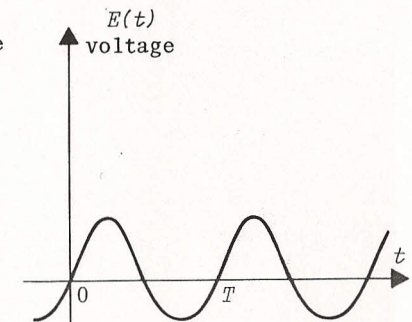


Figure 12.5

of closeness might be taken as the energy of the difference voltage. This is proportional to

$$\int_0^T [E(t) - q(t)]^2 dt,$$

which is again a least squares criterion.

3. Let $y(x)$ be the vertical displacement of a uniform flexible string whose equilibrium position is along the x -axis from $x=0$ to $x=l$. The elastic potential energy of the string is proportional to

$$\int_0^l [y(x)]^2 dx.$$

If $q(x)$ is to be an approximation to the displacement, then as before, the energy integral

$$\int_0^l [y(x) - q(x)]^2 dx$$

determines a least squares criterion for the closeness of the approximation.

Least squares approximation is also used in situations where there is no a priori justification for its use, such as for approximating business cycles, population growth curves, sales curves, and so forth. It is used in these cases because of its mathematical simplicity. In general, if no other error criterion is immediately apparent for an approximation problem, the least squares criterion is the one most often chosen.

In the next section we develop the mathematical theory of the least squares approximation of a function by a linear combination of sinusoidal functions.

GENERAL THEORY

Let $f(t)$ be a given continuous function defined over an interval of the t -axis. We first consider the case when the interval is

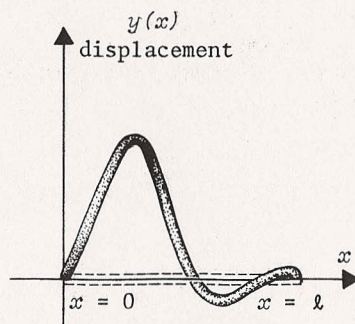


Figure 12.6

$[0, 2\pi]$ and later consider the general case when the interval is $[0, T]$ for arbitrary T . Analogous to Eq. (12.3) with $T=2\pi$, we desire to approximate $f(t)$ by a function of the form

$$g(t) = A_0 + A_1 \sin(t - \delta_1) + A_2 \sin 2(t - \delta_2) + \cdots + A_n \sin n(t - \delta_n)$$

for some fixed integer n . Since $\sin k(t - \delta_k)$ is expressible as a linear combination of $\sin kt$ and $\cos kt$, we can write $g(t)$ in the alternate form

$$g(t) = c_0 + c_1 \cos t + c_2 \cos 2t + \cdots + c_n \cos nt + d_1 \sin t + d_2 \sin 2t + \cdots + d_n \sin nt. \quad (12.5)$$

Such a function is called a *trigonometric polynomial of order n* . Our problem is to find values of $c_0, c_1, \dots, c_n, d_1, \dots, d_n$ such that $g(t)$ is the least squares approximation to $f(t)$ over the interval $[0, 2\pi]$. That is, the coefficients are to be chosen so that the least squares error

$$\int_0^{2\pi} [f(t) - g(t)]^2 dt \quad (12.6)$$

is as small as possible.

Since the integral in Eq. (12.6) is a function of the $2n+1$ coefficients $c_0, c_1, \dots, c_n, d_1, \dots, d_n$, it is possible to use calculus to find the minimum value of the least squares error and the corresponding values of these $2n+1$ coefficients. However, an approach using Linear Algebra will give us greater insight into the nature of the approximation process. Moreover, the method we discuss can be applied to many least square problems besides those in this text. We will need the following three facts:

1. The function $f(t)$ we are attempting to approximate may be viewed as a vector in the vector space $C[0, 2\pi]$ — the space of all continuous functions on $[0, 2\pi]$.
2. Since the approximating function $g(t)$ is a linear combination of $1, \cos t, \dots, \cos nt, \sin t, \dots, \sin nt$, we may view $g(t)$ as a vector in the subspace W of $C[0, 2\pi]$ spanned by these $2n+1$ vectors.

3. Since

$$\|f - g\| = \sqrt{\int_0^{2\pi} [f(t) - g(t)]^2 dt} \quad (12.7)$$

is the distance between $f(t)$ and $g(t)$ in the norm generated by the inner product

$$\langle u, v \rangle = \int_0^{2\pi} u(t)v(t) dt, \quad (12.8)$$

the least squares error

$$\int_0^{2\pi} [f(t) - g(t)]^2 dt \quad (12.9)$$

represents the square of the distance $\|f - g\|$.

In light of these remarks, the problem of finding a trigonometric polynomial $g(t)$ which minimizes the least squares error given by (12.9) is equivalent to the problem of finding a vector g in the subspace W which minimizes the distance $\|f - g\|$. The latter problem can be solved by use of the following theorem from the theory of inner product spaces (Fig. 12.7)

THEOREM 12.1 *Let f be a vector in an inner product space and let W be a finite dimensional subspace. Then the vector g in W which minimizes the distance $\|f - g\|$ is $\text{proj}_W f$, the orthogonal projection of f onto W . If the vectors*

$$g_0, g_1, \dots, g_m$$

form an orthonormal basis for W , then $\text{proj}_W f$ is given by

$$\text{proj}_W f = \langle f, g_0 \rangle g_0 + \langle f, g_1 \rangle g_1 + \dots + \langle f, g_m \rangle g_m \quad (12.10)$$

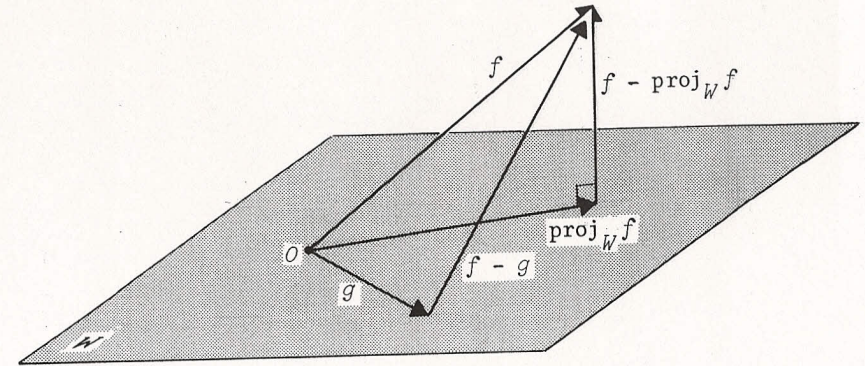


Figure 12.7

To apply this theorem, we must first find an orthonormal basis for the subspace W spanned by the $2n+1$ vectors $1, \cos t, \cos 2t, \dots, \cos nt, \sin t, \sin 2t, \dots, \sin nt$. A direct calculation (see Exercise 12.6) verifies that these $2n+1$ vectors are orthogonal relative to the inner product (12.8). Consequently, we need only divide each one of these vectors by its length to generate an orthonormal basis for W . The result is (see Exercise 12.7):

$$g_0 = \frac{1}{\sqrt{2\pi}}, g_1 = \frac{1}{\sqrt{\pi}} \cos t, \dots, g_n = \frac{1}{\sqrt{\pi}} \cos nt,$$

$$g_{n+1} = \frac{1}{\sqrt{\pi}} \sin t, \dots, g_{2n} = \frac{1}{\sqrt{\pi}} \sin nt.$$

The orthogonal projection of f onto W is then given by (12.10):

$$\begin{aligned} \text{proj}_W f = & \langle f, g_0 \rangle \frac{1}{\sqrt{2\pi}} + \langle f, g_1 \rangle \frac{1}{\sqrt{\pi}} \cos t + \dots + \langle f, g_n \rangle \frac{1}{\sqrt{\pi}} \cos nt \\ & + \langle f, g_{n+1} \rangle \frac{1}{\sqrt{\pi}} \sin t + \dots + \langle f, g_{2n} \rangle \frac{1}{\sqrt{\pi}} \sin nt. \end{aligned} \quad (12.11)$$

To simplify our notation, let us define

$$a_0 = 2\langle f, g_0 \rangle \frac{1}{\sqrt{2\pi}} = \frac{1}{\pi} \int_0^{2\pi} f(t) dt,$$

$$a_k = \langle f, g_k \rangle \frac{1}{\sqrt{\pi}} = \frac{1}{\pi} \int_0^{2\pi} f(t) \cos kt dt, \quad k = 1, 2, \dots, n$$

$$b_k = \langle f, g_{n+k} \rangle \frac{1}{\sqrt{\pi}} = \frac{1}{\pi} \int_0^{2\pi} f(t) \sin kt dt, \quad k = 1, 2, \dots, n.$$

Equation (12.11) can then be written as

$$\text{proj}_W f = \frac{1}{2}a_0 + a_1 \cos t + \dots + a_n \cos nt + b_1 \sin t + \dots + b_n \sin nt.$$

In summary, we have the following result:

THEOREM 12.2 If $f(t)$ is continuous on $[0, 2\pi]$, the trigonometric function $g(t)$ of the form

$$g(t) = \frac{1}{2}a_0 + a_1 \cos t + \dots + a_n \cos nt + b_1 \sin t + \dots + b_n \sin nt$$

which minimizes the least squares error

$$\int_0^{2\pi} [f(t) - g(t)]^2 dt$$

has coefficients

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \cos kt dt, \quad k = 0, 1, 2, \dots, n$$

$$b_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \sin kt dt, \quad k = 1, 2, \dots, n.$$

If the original function $f(t)$ is defined over the interval $[0, T]$ instead of $[0, 2\pi]$, a change of scale will yield the following result (see Exercise 12.8):

THEOREM 12.3 If $f(t)$ is continuous on $[0, T]$, the trigonometric function $g(t)$ of the form

$$g(t) = \frac{1}{2}a_0 + a_1 \cos \frac{2\pi}{T}t + \dots + a_n \cos \frac{2n\pi}{T}t \\ + b_1 \sin \frac{2\pi}{T}t + \dots + b_n \sin \frac{2n\pi}{T}t$$

which minimizes the least squares error

$$\int_0^T [f(t) - g(t)]^2 dt$$

has coefficients

$$a_k = \frac{2}{T} \int_0^T f(t) \cos \frac{2k\pi t}{T} dt, \quad k = 0, 1, 2, \dots, n$$

$$b_k = \frac{2}{T} \int_0^T f(t) \sin \frac{2k\pi t}{T} dt, \quad k = 1, 2, \dots, n.$$

EXAMPLE 12.1 Let a sound wave $p(t)$ have a saw-tooth pattern with a basic frequency of 5000 cps (Fig. 12.8). Units have been chosen so that the normal atmospheric pressure is at the zero level, and the maximum amplitude of the wave is A . The basic period of the wave is $T = 1/5000 = .0002$ seconds. From $t = 0$ to $t = T$, $p(t)$ has the equation

$$p(t) = \frac{2A}{T} \left(\frac{T}{2} - t \right).$$

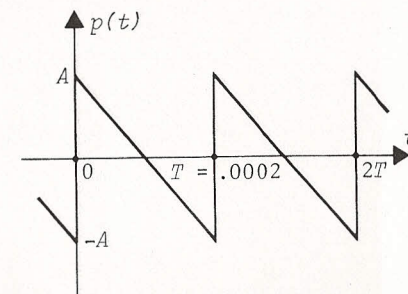


Figure 12.8

Theorem 12.3 then yields the following (verify):

$$a_0 = \frac{2}{T} \int_0^T p(t) dt = \frac{2}{T} \int_0^T \frac{2A}{T} \left(\frac{T}{2} - t \right) dt = 0,$$

$$a_k = \frac{2}{T} \int_0^T p(t) \cos \frac{2k\pi t}{T} dt = \frac{2}{T} \int_0^T \frac{2A}{T} \left(\frac{T}{2} - t \right) \cos \frac{2k\pi t}{T} dt = 0, \quad k = 1, 2, \dots$$

$$b_k = \frac{2}{T} \int_0^T p(t) \sin \frac{2k\pi t}{T} dt = \frac{2}{T} \int_0^T \frac{2A}{T} \left(\frac{T}{2} - t \right) \sin \frac{2k\pi t}{T} dt = \frac{2A}{k\pi}, \quad k = 1, 2, \dots$$

Let us investigate how the sound wave $p(t)$ is perceived by the human ear. We notice that $4/T = 20,000$ cps, so that we need only go up to $k = 4$ in the above formulas. The least squares approximation to $p(t)$ is then

$$q(t) = \frac{2A}{\pi} \left[\sin \frac{2\pi}{T} t + \frac{1}{2} \sin \frac{4\pi}{T} t + \frac{1}{3} \sin \frac{6\pi}{T} t + \frac{1}{4} \sin \frac{8\pi}{T} t \right].$$

The four sinusoidal terms have frequencies of 5,000, 10,000, 15,000, and 20,000 cps, respectively. In Fig. 12.9 we have plotted $p(t)$ and $q(t)$ over one period. Although $q(t)$ is not a very good point-by-point approximation to $p(t)$, to the ear, both $p(t)$ and $q(t)$ produce the same sensation of sound.

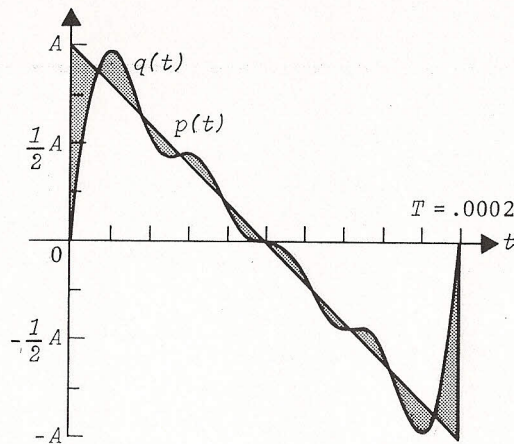


Figure 12.9

As might be expected, the least squares approximation becomes better as the number of terms in the approximating trigonometric polynomial becomes larger. In more advanced courses, it is shown that the least squares error tends to zero as n approaches infinity. For a function $f(t)$ defined over the interval $[0, 2\pi]$, this limiting approximation is denoted by

$$f(t) = \frac{1}{2} a_0 + \sum_{k=1}^{\infty} (a_k \cos kt + b_k \sin kt) \quad (12.12)$$

and is called the *Fourier series* of $f(t)$ on the interval $[0, 2\pi]$. The equality in this equation denotes an equality between the two sides of the equation considered as vectors in $C[0, 2\pi]$. To be precise, Eq. (12.12) denotes the fact that the quantity

$$\int_0^{2\pi} \left[f(t) - \frac{1}{2} a_0 - \sum_{k=1}^n (a_k \cos kt + b_k \sin kt) \right]^2 dt$$

tends to zero as n approaches infinity. Whether the Fourier series of $f(t)$ converges to $f(t)$ for each t is another question, and a more difficult one. For most continuous functions encountered in applications, the Fourier series does indeed converge to its corresponding function for each value of t .

EXERCISES

- 12.1 Find the trigonometric polynomial of order three which is the least squares approximation to the function $f(t) = (t - \pi)^2$ over the interval $[0, 2\pi]$.
- 12.2 Find the trigonometric polynomial of order four which is the least squares approximation to the function $f(t) = t^2$ over the interval $[0, T]$.
- 12.3 Find the trigonometric polynomial of order four which is the least squares approximation to the function $f(t)$ over the interval $[0, 2\pi]$ where

$$f(t) = \begin{cases} \sin t & 0 \leq t \leq \pi \\ 0 & \pi < t \leq 2\pi. \end{cases}$$

- 12.4 Find the trigonometric polynomial of arbitrary order n which is the least squares approximation to the function $f(t) = \sin \frac{1}{2}t$ over the interval $[0, 2\pi]$.
- 12.5 Find the trigonometric polynomial of arbitrary order n which is the least squares approximation to the function $f(t)$ over the interval $[0, T]$ where

$$f(t) = \begin{cases} t & 0 \leq t \leq \frac{1}{2}T \\ T - t & \frac{1}{2}T < t \leq T \end{cases}$$

- 12.6 Show that the $2n+1$ functions

$$1, \cos t, \cos 2t, \dots, \cos nt, \sin t, \sin 2t, \dots, \sin nt$$

are orthogonal over the interval $[0, 2\pi]$ relative to the inner product $\langle u, v \rangle$ defined by Eq. (12.8).

- 12.7 For the distance formula defined in (12.7), show that

$$\begin{aligned} \text{(a)} \quad \|1\| &= 1/\sqrt{2\pi} \\ \|\cos kt\| &= 1/\sqrt{\pi} \text{ for } k = 1, 2, \dots \\ \|\sin kt\| &= 1/\sqrt{\pi} \text{ for } k = 1, 2, \dots \end{aligned}$$

- 12.8 If $f(t)$ is defined and continuous on the interval $[0, T]$, show that $f(2\pi\tau/T)$ is defined and continuous for τ in the interval $[0, 2\pi]$. Use this fact to show how Theorem 12.3 follows from Theorem 12.2.

Linear Programming 1: A Geometric Approach

A geometric technique for maximizing or minimizing a linear expression in two variables subject to a set of linear constraints is described.

PREREQUISITES: Linear systems
 Linear inequalities

INTRODUCTION

The study of linear programming theory has expanded greatly since the pioneer work of George Dantzig in the late nineteen-forties. Today, linear programming is applied to a wide variety of problems in industry and science. In this chapter we present a geometric approach to the solution of simple linear programming problems. In Chapters 14 and 15 we develop the algebraic theory required to solve more general problems in this field.

Let us begin with some examples:

EXAMPLE 13.1 A candy manufacturer has 130 pounds of chocolate-covered cherries and 170 pounds of chocolate-covered mints in stock. He decides to sell them in the form of two different mixtures. One

mixture will contain half cherries and half mints and will sell for \$2.00 per pound. The other mixture will contain one-third cherries and two-thirds mints and will sell for \$1.25 per pound. How many pounds of each mixture should the candy manufacturer prepare in order to maximize his sales revenue?

Let us first formulate this problem mathematically. Let the mixture of half cherries and half mints be called mix A , and let x_1 be the number of pounds of this mixture to be prepared. Let the mixture of one-third cherries and two-thirds mints be called mix B , and let x_2 be the number of pounds of this mixture to be prepared. Since mix A sells for \$2.00 per pound and mix B sells for \$1.25 per pound, the total sales z (in dollars) will be

$$z = 2.00x_1 + 1.25x_2.$$

Since each pound of mix A contains $1/2$ pound of cherries and each pound of mix B contains $1/3$ pound of cherries, the total number of pounds of cherries used in both mixtures is

$$\frac{1}{2}x_1 + \frac{1}{3}x_2.$$

Similarly, since each pound of mix A contains $1/2$ pound of mints and each pound of mix B contains $2/3$ pound of mints, the total number of pounds of mints used in both mixtures is

$$\frac{1}{2}x_1 + \frac{2}{3}x_2.$$

Because the manufacturer can use at most 130 pounds of cherries and 170 pounds of mints, we must have

$$\frac{1}{2}x_1 + \frac{1}{3}x_2 \leq 130$$

$$\frac{1}{2}x_1 + \frac{2}{3}x_2 \leq 170.$$

Also, since x_1 and x_2 cannot be negative numbers, we must have

$$x_1 \geq 0 \quad \text{and} \quad x_2 \geq 0.$$

The problem can therefore be formulated mathematically as follows:

Find values of x_1 and x_2 which maximize

$$z = 2.00x_1 + 1.25x_2$$

subject to

$$\frac{1}{2}x_1 + \frac{1}{3}x_2 \leq 130$$

$$\frac{1}{2}x_1 + \frac{2}{3}x_2 \leq 170$$

$$x_1 \geq 0$$

$$x_2 \geq 0.$$

In the next section we shall show how to solve this type of mathematical problem geometrically.

EXAMPLE 13.2 A woman has up to \$10,000 to invest. Her broker suggests investing in two bonds, A and B . Bond A is a rather risky bond with an annual yield of 10%, and bond B is a rather safe bond with an annual yield of 7%. After some consideration, she decides to invest at most \$6,000 in bond A , at least \$2,000 in bond B , and to invest at least as much in bond A as in bond B . How should she invest her \$10,000 in order to maximize her annual yield?

In order to formulate this problem mathematically, let x_1 be the number of dollars to be invested in bond A and let x_2 be the number of dollars to be invested in bond B . Since each dollar invested in bond A earns \$.10 per year and each dollar invested in bond B earns \$.07 per year, the total dollar amount z earned each year by both bonds is

$$z = .10x_1 + .07x_2.$$

The constraints imposed can be formulated mathematically as follows:

$$\text{Invest no more than } \$10,000: \quad x_1 + x_2 \leq 10,000$$

$$\text{Invest at most } \$6,000 \text{ in bond } A: \quad x_1 \leq 6,000$$

$$\text{Invest at least } \$2,000 \text{ in bond } B: \quad x_2 \geq 2,000$$

$$\text{Invest at least as much in bond } A \text{ as in bond } B: \quad x_1 \geq x_2.$$

We also have the implicit assumption that x_1 and x_2 are nonnegative:

$$x_1 \geq 0 \quad \text{and} \quad x_2 \geq 0.$$

Thus, the complete mathematical formulation of the problem is as follows:

Find values of x_1 and x_2 which maximize

$$z = .10x_1 + .07x_2$$

subject to

$$x_1 + x_2 \leq 10,000$$

$$x_1 \leq 6,000$$

$$x_2 \geq 2,000$$

$$x_1 - x_2 \geq 0$$

$$x_1 \geq 0$$

$$x_2 \geq 0.$$

EXAMPLE 13.3 A student desires to design a breakfast of corn flakes and milk which is as economical as possible. On the basis of what he eats during his other meals, he decides that his breakfast should supply him with at least nine grams of protein, at least one-third the recommended daily allowance (RDA) of vitamin D, and at least one-fourth the RDA of calcium. He finds the following nutrition information on the milk and corn flakes containers:

	Milk ½ cup	Corn Flakes 1 ounce
Cost	7.5¢	5.0¢
Protein	4 grams	2 grams
Vitamin D	1/8 of RDA	1/10 of RDA
Calcium	1/6 of RDA	none

In order not to have his mixture too soggy or too dry, the student decides to limit himself to mixtures which contain one to three ounces of corn flakes per cup of milk, inclusive. What quantities of milk and corn flakes should he use to minimize the cost of his breakfast?

For the mathematical formulation of this problem, let x_1 be the quantity of milk used measured in ½-cup units and let x_2 be the quantity of corn flakes used measured in 1-ounce units. Then if z is the cost of the breakfast in cents, we may write the following:

Cost of breakfast: $z = 7.5x_1 + 5.0x_2$

At least nine grams of protein: $4x_1 + 2x_2 \geq 9$

At least 1/3 of RDA of vitamin D: $\frac{1}{8}x_1 + \frac{1}{10}x_2 \geq \frac{1}{3}$

At least 1/4 of RDA of calcium: $\frac{1}{6}x_1 \geq \frac{1}{4}$

At least one ounce of corn flakes per cup (2 ½-cups) of milk: $x_2/x_1 \geq \frac{1}{2}$ (or $x_1 - 2x_2 \leq 0$)

At most three ounces of corn flakes per cup (2 ½-cups) of milk: $x_2/x_1 \leq \frac{3}{2}$ (or $3x_1 - 2x_2 \geq 0$)

As before, we also have the implicit assumption that $x_1 \geq 0$ and $x_2 \geq 0$. Thus the complete mathematical formulation of the problem is as follows

Find values of x_1 and x_2 which minimize

$$z = 7.5x_1 + 5.0x_2$$

subject to

$$4x_1 + 2x_2 \geq 9$$

$$\frac{1}{8}x_1 + \frac{1}{10}x_2 \geq \frac{1}{3}$$

$$\frac{1}{6}x_1 \geq \frac{1}{4}$$

$$x_1 - 2x_2 \leq 0$$

$$3x_1 - 2x_2 \geq 0$$

$$x_1 \geq 0$$

$$x_2 \geq 0.$$

GEOMETRIC SOLUTION OF LINEAR PROGRAMMING PROBLEMS

Each of the three examples in the introduction is a special case of the following problem:

PROBLEM 13.1 Find values of x_1 and x_2 which either maximize or minimize

$$z = c_1x_1 + c_2x_2 \quad (13.1)$$

subject to

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 & (\leq)(\geq)(=) b_1 \\ a_{21}x_1 + a_{22}x_2 & (\leq)(\geq)(=) b_2 \\ & \vdots \\ a_{m1}x_1 + a_{m2}x_2 & (\leq)(\geq)(=) b_m \end{aligned} \quad (13.2)$$

and

$$x_1 \geq 0, \quad x_2 \geq 0. \quad (13.3)$$

In Eqs. (13.2), only one of the symbols $\leq, \geq, =$ is to be used in each of the m equations.

Problem 13.1 is the *general linear programming problem* in two variables. The linear function z in (13.1) is called the *objective function*. Equations (13.2) and (13.3) are called the *constraints*; in particular, Eqs. (13.3) are called the *nonnegativity constraints* on the variables x_1 and x_2 .

We shall now show how to solve a linear programming problem in two variables graphically. A pair of values (x_1, x_2) which satisfy all of the constraints is called a *feasible solution*. The set of all feasible solutions determines a subset of the x_1x_2 -plane called the *feasible region*. Our desire is to find a feasible solution which maximizes the objective function. Such a solution is called an *optimal solution*.

To examine the feasible region of a linear programming problem, let us note that each constraint of the form

$$a_{i1}x_1 + a_{i2}x_2 = b_i$$

defines a line in the x_1x_2 -plane, while each constraint of the form

$$a_{i1}x_1 + a_{i2}x_2 \leq b_i$$

or

$$a_{i1}x_1 + a_{i2}x_2 \geq b_i$$

defines a half-plane which includes its boundary line

$$a_{i1}x_1 + a_{i2}x_2 = b_i.$$

Thus, the feasible region is always an intersection of finitely many lines and half-planes. For example, the four constraints

$$\frac{1}{2}x_1 + \frac{1}{3}x_2 \leq 130$$

$$\frac{1}{2}x_1 + \frac{2}{3}x_2 \leq 170$$

$$x_1 \geq 0$$

$$x_2 \geq 0$$

of Example 13.1 define the half-planes illustrated in Fig. 13.1(a), (b), (c), and (d). The feasible region of this problem is thus the intersection of these four half-planes, which is illustrated in Fig. 13.1(e).

It can be shown that the feasible region of a linear programming problem has a boundary consisting of a finite number of straight-line segments. If the feasible region can be enclosed in a sufficiently large circle, it is called *bounded* (Fig. 13.1(e)); otherwise it is called *unbounded* (Fig. 13.5). If the feasible region is *empty* (contains no points), then the constraints are inconsistent and the linear programming problem has no solution (Fig. 13.6).

Those boundary points of a feasible region which are intersections of two of the straight-line boundary segments are called *extreme points*. (They are also called *corner points* or *vertex points*.) For example, from Fig. 13.1(e), the feasible region of Example 13.1 has four extreme points:

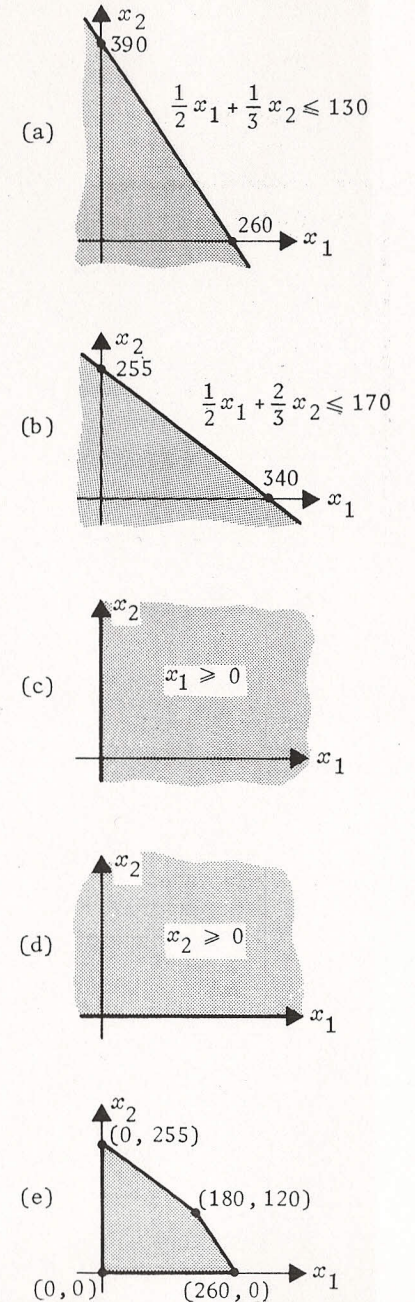


Figure 13.1

$$(0, 0), \quad (0, 255), \quad (180, 120), \quad (260, 0). \quad (13.4)$$

The importance of the extreme points of a feasible region is shown by the following theorem:

THEOREM 13.1 *If the feasible region of a linear programming problem is nonempty and bounded, then the objective function attains both a maximum and minimum value and these occur at extreme points of the feasible region. If the feasible region is unbounded then the objective function may or may not attain a maximum or minimum value; however, if it attains a maximum or minimum value, it does so at an extreme point.*

Figure 13.2 suggests the idea behind the proof of this theorem. Since the objective function

$$z = c_1x_1 + c_2x_2$$

of a linear programming problem is a linear function of x_1 and x_2 , its level curves (the curves along which z has constant values) are straight lines. As we move in a direction perpendicular to these level curves, the objective function either increases or decreases monotonically. Within a bounded feasible region, the maximum and minimum values of z must therefore occur at extreme points, as Figure 13.2 indicates.

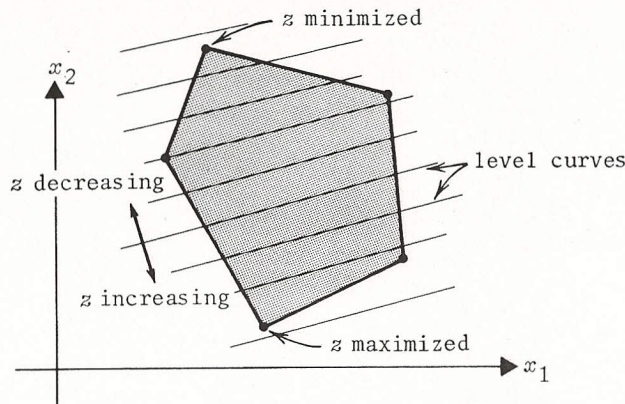


Figure 13.2

In the next few examples, we use Theorem 13.1 to solve several linear programming problems and illustrate the variations in the nature of the solutions which may occur.

EXAMPLE 13.1 (REVISITED) From Figure 13.1(e), we see that the feasible region of Example 13.1 is bounded. Consequently, from Theorem 13.1 the objective function

$$z = 2.00x_1 + 1.25x_2$$

attains both its minimum and maximum values at extreme points. The four extreme points and the corresponding values of z are given in the following table

Extreme point (x_1, x_2)	Value of $z = 2.00x_1 + 1.25x_2$
(0, 0)	0
(0, 255)	318.75
(180, 120)	510.00
(260, 0)	520.00

We see that the largest value of z is 520.00, and the corresponding optimal solution is (260, 0). Thus, the candy manufacturer attains maximum sales of \$520 when he produces 260 pounds of mixture A and none of mixture B.

EXAMPLE 13.4 Find values of x_1 and x_2 which maximize

$$z = x_1 + 3x_2$$

subject to

$$2x_1 + 3x_2 \leq 24$$

$$x_1 - x_2 \leq 7$$

$$x_2 \leq 6$$

$$x_1 \geq 0$$

$$x_2 \geq 0.$$

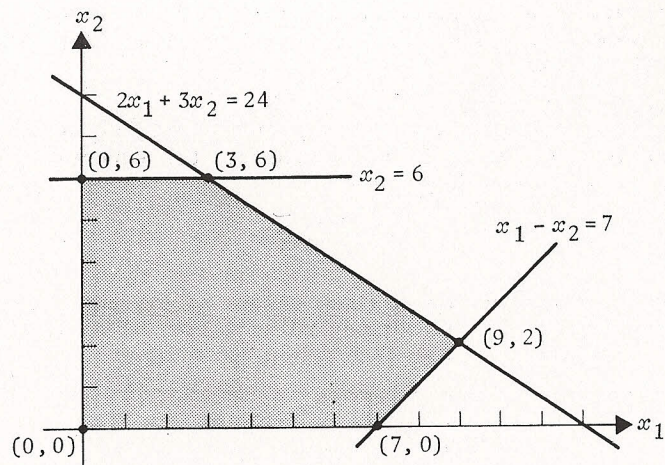


Figure 13.3

SOLUTION In Fig. 13.3 we have drawn the feasible region of this problem. Since it is bounded, the maximum value of z is attained at one of the five extreme points. The values of the objective function at the five extreme points are given in the following table:

Extreme point (x_1, x_2)	Value of $z = x_1 + 3x_2$
(0, 6)	18
(3, 6)	21
(9, 2)	15
(7, 0)	7
(0, 0)	0

From this table, the maximum value of z is 21, which is attained at $x_1 = 3$ and $x_2 = 6$.

EXAMPLE 13.5 Find values of x_1 and x_2 which maximize

$$z = 4x_1 + 6x_2$$

subject to

$$2x_1 + 3x_2 \leq 24$$

$$x_1 - x_2 \leq 7$$

$$x_2 \leq 6$$

$$x_1 \geq 0$$

$$x_2 \geq 0.$$

SOLUTION The constraints in this problem are identical to the constraints in Example 13.4, and so the feasible region of this problem is also given by Fig. 13.3. The values of the objective function at the extreme points are as follows:

Extreme point (x_1, x_2)	Value of $z = 4x_1 + 6x_2$
(0, 6)	36
(3, 6)	48
(9, 2)	48
(7, 0)	28
(0, 0)	0

We see that the objective function attains a maximum value of 48 at two adjacent extreme points, (3, 6) and (9, 2). This shows that an optimal solution to a linear programming problem need not be unique. As we ask the reader to show in Exercise 13.9, if the objective function has the same value at two adjacent extreme points, it has the same value at all points on the straight-line boundary segment connecting the two extreme points. Thus, in this example the maximum value of z is attained at all points on the straight-line segment connecting the extreme points (3, 6) and (9, 2).

EXAMPLE 13.6 Find values of x_1 and x_2 which minimize

$$z = 2x_1 - x_2$$

subject to

$$2x_1 + 3x_2 = 12$$

$$2x_1 - 3x_2 \geq 0$$

$$x_1 \geq 0$$

$$x_2 \geq 0.$$

SOLUTION In Fig. 13.4 we have drawn the feasible region of this problem. Because one of the constraints is an equality constraint, the feasible region is a straight line segment with two extreme points. The values of z at the two extreme points are as follows:

Extreme point (x_1, x_2)	Value of $z = 2x_1 - x_2$
(3, 2)	4
(6, 0)	12

The minimum value of z is thus 4, and is attained at $x_1 = 3$ and $x_2 = 2$.

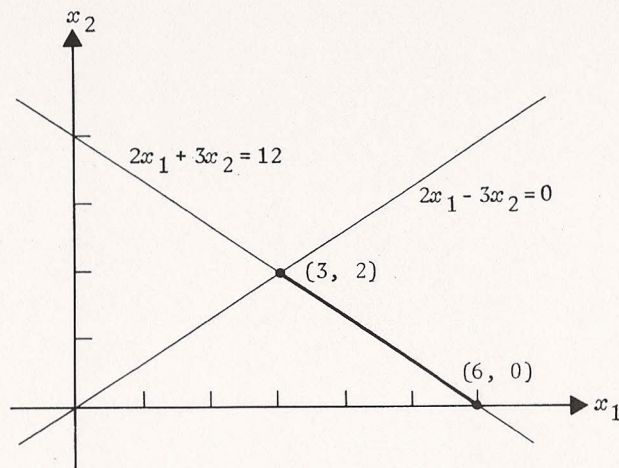


Figure 13.4

EXAMPLE 13.7 Find values of x_1 and x_2 which maximize

$$z = 2x_1 + 5x_2$$

subject to

$$2x_1 + x_2 \geq 8$$

$$-4x_1 + x_2 \leq 2$$

$$2x_1 - 3x_2 \leq 0$$

$$x_1 \geq 0$$

$$x_2 \geq 0.$$

SOLUTION The feasible region of this linear programming problem is illustrated in Fig. 13.5. Since it is unbounded, we are not assured by Theorem 13.1 that the objective function attains a maximum value. In fact, it is easily seen that since the feasible region contains points for which both x_1 and x_2 are arbitrarily large and positive, then the objective function

$$z = 2x_1 + 5x_2$$

can be made arbitrarily large and positive. This problem has no optimal solution. Instead, we say the problem has an *unbounded solution*.

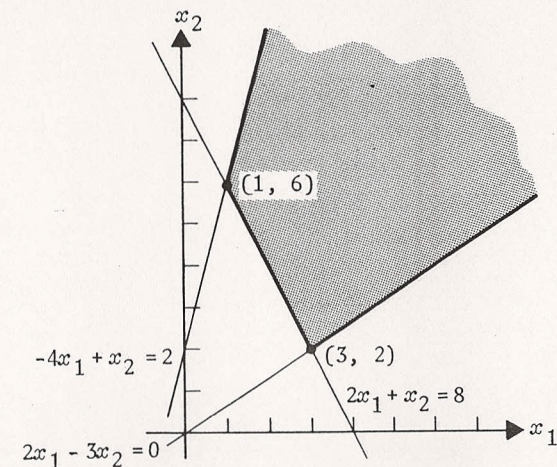


Figure 13.5

EXAMPLE 13.8 Find values of x_1 and x_2 which maximize

$$z = -5x_1 + x_2$$

subject to

$$2x_1 + x_2 \geq 8$$

$$-4x_1 + x_2 \leq 2$$

$$2x_1 - 3x_2 \leq 0$$

$$x_1 \geq 0$$

$$x_2 \geq 0.$$

SOLUTION The above constraints are the same as those in Example 13.7, so that the feasible region of this problem is also given by Fig. 13.5. In Exercise 13.10, we ask the reader to show that the objective function of this problem attains a maximum within the feasible region. By Theorem 13.1, this maximum must be attained at an extreme point. The values of z at the two extreme points of the feasible region are given by

Extreme point (x_1, x_2)	Value of $z = -5x_1 + x_2$
(1, 6)	1
(3, 2)	-13

The maximum value of z is thus 1, and is attained at the extreme point $x_1 = 1, x_2 = 6$.

EXAMPLE 13.9 Find the values x_1 and x_2 which minimize

$$z = 3x_1 - 8x_2$$

subject to

$$2x_1 - x_2 \leq 4$$

$$3x_1 + 11x_2 \leq 33$$

$$3x_1 + 4x_2 \geq 24$$

$$x_1 \geq 0$$

$$x_2 \geq 0.$$

SOLUTION As can be seen from Fig. 13.6, the intersection of the five half-planes defined by the five constraints is empty. This linear programming problem has no feasible solutions since the constraints are inconsistent.

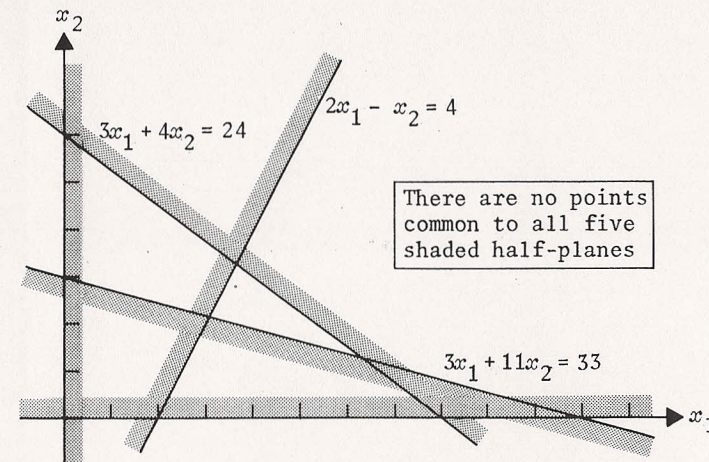


Figure 13.6

EXERCISES

13.1 Find values of x_1 and x_2 which maximize

$$z = 3x_1 + 2x_2$$

subject to

$$\begin{aligned}
 2x_1 + 3x_2 &\leq 6 \\
 2x_1 - x_2 &\geq 0 \\
 x_1 &\leq 2 \\
 x_2 &\leq 1 \\
 x_1 &\geq 0 \\
 x_2 &\geq 0.
 \end{aligned}$$

13.2 Find values of x_1 and x_2 which minimize

$$\begin{aligned}
 z &= 3x_1 - 5x_2 \\
 \text{subject to} \\
 2x_1 - x_2 &\leq -2 \\
 4x_1 - x_2 &\geq 0 \\
 x_2 &\leq 3 \\
 x_1 &\geq 0 \\
 x_2 &\geq 0.
 \end{aligned}$$

13.3 Find values of x_1 and x_2 which minimize

$$\begin{aligned}
 z &= -3x_1 + 2x_2 \\
 \text{subject to} \\
 3x_1 - x_2 &\geq -5 \\
 -x_1 + x_2 &\geq 1 \\
 2x_1 + 4x_2 &\geq 12 \\
 x_1 &\geq 0 \\
 x_2 &\geq 0.
 \end{aligned}$$

13.4 Solve the linear programming problem posed in Example 13.2.

13.5 Solve the linear programming problem posed in Example 13.3.

13.6 A trucking firm ships the containers of two companies, A and B . Each container of Company A weighs 40 pounds and is 2 cubic feet in volume. Each container of Company B weighs 50 pounds and is 3 cubic feet in volume. The trucking firm charges Company A \$2.20 for each container shipped, and charges Company B \$3.00 for each container shipped. If one of the firm's trucks cannot carry more than 37,000 pounds and cannot hold more than 2000 cubic feet, how many containers from companies A and B should a truck carry to maximize the shipping charges.

13.7 Repeat Exercise 13.7 if the trucking firm raises its price for shipping a container of Company A to \$2.50.

13.8 A manufacturer produces sacks of chicken feed from two ingredients, A and B . Each sack is to contain at least 10 ounces of nutrient N_1 , at least 8 ounces of nutrient N_2 , and at least 12 ounces of nutrient N_3 . Each pound of ingredient A contains 2 ounces of nutrient N_1 , 2 ounces of nutrient N_2 , and 6 ounces of nutrient N_3 . Each pound of ingredient B contains 5 ounces of nutrient N_1 , 3 ounces of nutrient N_2 , and 4 ounces of nutrient N_3 . If ingredient A costs 8¢ per pound and ingredient B costs 9¢ per pound, how much of each ingredient should the manufacturer use in each sack of feed to minimize his costs?

13.9 If the objective function of a linear programming problem has the same value at two adjacent extreme points, show that it has the same value at all points on the straight-line segment connecting the two extreme points. Hint: If (x_1', x_2') and (x_1'', x_2'') are any two points in the plane, a point (x_1, x_2) lies on the straight-line segment connecting them if

$$x_1 = tx_1' + (1-t)x_1''$$

and

$$x_2 = tx_2' + (1-t)x_2''$$

where t is a number in the interval $[0, 1]$.

13.10 Show that the objective function in Example 13.8 attains a maximum value in the feasible set. Hint: Examine the level curves of the objective function.

Linear Programming 2: Basic Concepts

The basic concepts needed to develop the simplex method for solving linear programming problems are presented.

PREREQUISITES: Linear Systems
 Matrices
 Linear independence
 Euclidean space R^n
 Chapter 13: Linear Programming I

INTRODUCTION

In the last chapter we presented a geometric technique for solving linear programming problems in two variables. However, this technique is not practical for the solution of linear programming problems in three or more variables. In this chapter we develop the fundamental ideas behind an algebraic technique, called the *simplex method*, for solving linear programming problems in any number of variables. The simplex method itself is presented in Chapter 15.

The *general linear programming problem in n variables*, described below, is analogous to Problem 13.1 in the last chapter.

PROBLEM 14.1 Find values of x_1, x_2, \dots, x_n which either maximize or minimize

$$z = c_1x_1 + c_2x_2 + \dots + c_nx_n \quad (14.1)$$

subject to

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n & (\leq) (\geq) (=) b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n & (\leq) (\geq) (=) b_2 \\ & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n & (\leq) (\geq) (=) b_m \end{aligned} \quad (14.2)$$

and

$$x_i \geq 0, \quad \text{for } i = 1, 2, \dots, n. \quad (14.3)$$

As in Chapter 13, the linear function z in (14.1) is called the *objective function* and conditions (14.2) and (14.3) are called the *constraints* of the problem.

For our purposes, it is necessary that we first convert our general linear programming problem to the following *standard form*:

PROBLEM 14.2 Find values of x_1, x_2, \dots, x_n which maximize

$$z = c_1x_1 + c_2x_2 + \dots + c_nx_n \quad (14.4)$$

subject to

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n & = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n & = b_2 \\ & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n & = b_m \end{aligned} \quad (14.5)$$

and

$$x_i \geq 0, \quad \text{for } i = 1, 2, \dots, n. \quad (14.6)$$

Any linear programming problem can always be put in this standard form using the following three steps:

Step 1. Convert a minimization problem to a maximization problem by defining a new objective function

$$z' = -z.$$

For example, the problem of minimizing the objective function

$$z = 2x_1 + 3x_2 - 5x_3$$

is equivalent to the problem of maximizing the objective function

$$z' = -2x_1 - 3x_2 + 5x_3.$$

Step 2. Convert a \geq constraint to a \leq constraint by multiplying the inequality by -1 . Thus, the constraint

$$-x_1 + 2x_2 - 4x_3 \geq 6$$

is the same as the constraint

$$x_1 - 2x_2 + 4x_3 \leq -6.$$

Step 3. Convert a \leq constraint to an equality constraint by adding a nonnegative *slack variable* to the lefthand side of the inequality. For example, if the original problem contains three variables, and one of the constraints is

$$x_1 - 2x_2 + 4x_3 \leq -6,$$

we add a new variable $x_4 \geq 0$ to the lefthand side to obtain

$$x_1 - 2x_2 + 4x_3 + x_4 = -6.$$

The variable x_4 takes up the slack between the two sides of the inequality. In this way, a new variable is added to the problem for each \leq constraint. We assign each slack variable introduced a coefficient $c_i = 0$ in the objective function, so that the objective function is not affected by the values of the slack variables.

EXAMPLE 14.1 Convert the following linear programming problem to one in standard form:

$$\text{Minimize } z = 3x_1 - 2x_2 + x_3 - x_4$$

subject to

$$2x_1 + 5x_2 - 6x_3 - x_4 \leq 2$$

$$x_1 - 7x_2 - 5x_3 + 2x_4 \geq 6$$

$$2x_1 - 8x_2 - 8x_3 + 6x_4 = 5$$

and

$$x_1, x_2, x_3, x_4 \geq 0.$$

SOLUTION The first step is to multiply the objective function by -1 to convert the problem to a maximization problem. The second step is to multiply the second constraint by -1 to convert it to a \leq inequality. The third step is to add slack variables x_5 and x_6 to the first and second constraints to convert them to equalities. The final problem in standard form is

$$\text{Maximize } z' = -3x_1 + 2x_2 - x_3 + x_4 + 0x_5 + 0x_6$$

subject to

$$2x_1 + 5x_2 - 6x_3 - x_4 + x_5 = 2$$

$$-x_1 + 7x_2 + 5x_3 - 2x_4 + x_6 = -6$$

$$2x_1 - 8x_2 - 8x_3 + 6x_4 = 5$$

and

$$x_1, x_2, x_3, x_4, x_5, x_6 \geq 0.$$

It will be convenient to use matrix notation to express Problem 14.2 in a more compact form as follows (the expression $x \geq 0$ below denotes that each entry of the vector x is nonnegative):

PROBLEM 14.2 (in matrix notation)

$$\text{Maximize } z = c^t x$$

subject to

$$Ax = b$$

and

$$x \geq 0,$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \quad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

In this formulation, the problem is to find a nonnegative vector x in the vector space R^n which satisfies the constraint condition $Ax = b$ and which makes the objective function $z = c^t x$ as large as possible. Analogous to the terminology in Chapter 13, any nonnegative vector x which satisfies the constraint $Ax = b$ is called a *feasible solution* to the linear programming problem. The set of all feasible solutions in R^n is called the *feasible set* or the *feasible region* of the problem. A feasible solution which maximizes the objective function is called an *optimal solution*. As in Chapter 13, there are three possible outcomes to a linear programming problem:

- (i) The constraints are inconsistent so that there are no feasible solutions.
- (ii) The feasible set is unbounded and the objective function can be made arbitrarily large.
- (iii) There is at least one optimal solution.

In most realistic applications only case (iii) arises.

We now examine the nature of the feasible set of a linear programming problem. To begin, let us introduce the following definition:

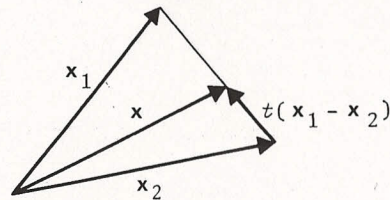
DEFINITION 14.1 A set of vectors in R^n is called **convex** if whenever x_1 and x_2 belong to the set, so does the vector

$$x = tx_1 + (1-t)x_2$$

for any number t in the interval $[0, 1]$.

Geometrically, the vector $x = tx_1 + (1-t)x_2$ lies on the line segment connecting the tips of the vectors x_1 and x_2 (Fig. 14.1). Thus, a convex set can be viewed as one in which the line segment connecting any two points in the set also belongs to the set. Figures 14.2(a), (b), and (c) illustrate three convex sets in R^2 . Figure 14.1(d) is an example of a set in R^2 which is not convex. In Exercise 14.7, we ask the reader to prove the following theorem:

THEOREM 14.1 *The feasible set of a standard linear programming problem is convex.*



$$x = x_2 + t(x_1 - x_2) = tx_1 + (1-t)x_2$$

Figure 14.1

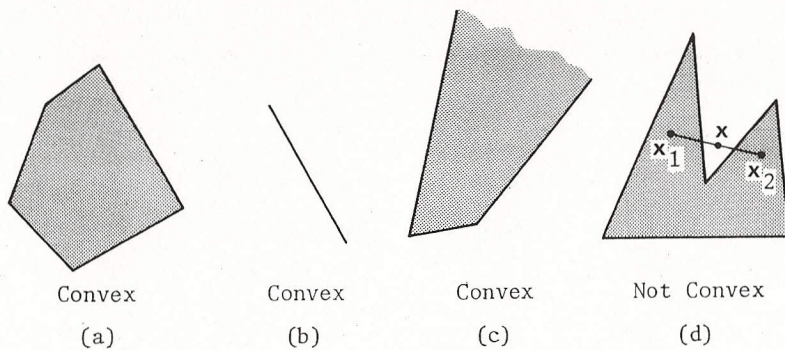


Figure 14.2

EXAMPLE 14.2 For the following linear programming problem:

$$\text{Maximize } z = 3x_1 + x_2 - 2x_3$$

subject to

$$3x_1 + x_2 + x_3 = 10$$

$$2x_1 - x_2 + 2x_3 = 10$$

and

$$x_1, x_2, x_3 \geq 0,$$

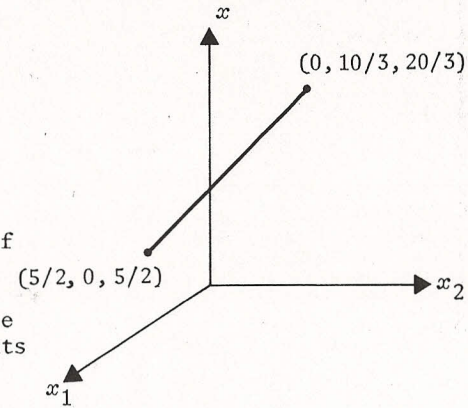
the feasible set consists of the portion of the intersection of the two planes

$$3x_1 + x_2 + x_3 = 10$$

$$2x_1 - x_2 + 2x_3 = 10$$

which lies in the first octant of $x_1x_2x_3$ -space. As we ask the reader to show in Exercise 14.8, this intersection consists of the line segment connecting the points

$(5/2, 0, 5/2)$ and $(0, 10/3, 20/3)$



(Fig. 14.3), which is clearly convex.

Figure 14.3

EXAMPLE 14.3 For the following linear programming problem:

$$\text{Maximize } z = x_1 - 3x_2 + 2x_3$$

subject to

$$2x_1 + 4x_2 + 3x_3 = 12$$

and

$$x_1, x_2, x_3 \geq 0,$$

the feasible set consists of the portion of the plane

$$2x_1 + 4x_2 + 3x_3 = 12$$

which lies in the first octant of $x_1x_2x_3$ -space. The reader can easily verify that Fig. 14.4 is a diagram of this feasible set and that it is convex.

In Chapter 13, we introduced the concept of an extreme point of a feasible set in x_1x_2 -space. For an arbitrary convex set in R^n , the corresponding definition is the following:

DEFINITION 14.2 A vector x in a convex set is an extreme point of the convex set if

$$x \neq \frac{1}{2}(x_1 + x_2)$$

for any two vectors x_1 and x_2 in the set

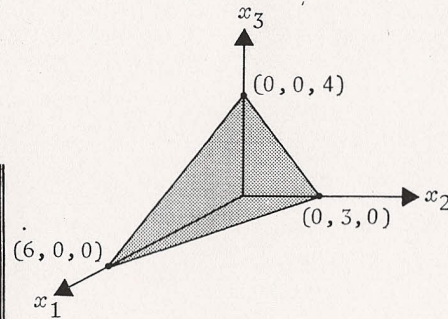


Figure 14.4

In other words, the extreme points of a convex set are those points which do not lie midway between any two points in the set. For example, the extreme points of the convex set illustrated in Fig. 14.2 are the two endpoints

$$(5/2, 0, 5/2) \text{ and } (0, 10/3, 20/3)$$

of the straight-line segment. For the convex set illustrated in Fig. 14.4, the extreme points are the three corner points

$$(6, 0, 0), (0, 3, 0), \text{ and } (0, 0, 4)$$

of the triangular-shaped region.

Recall that in the last chapter we called a set in R^2 bounded if it can be enclosed in a sufficiently large circle; that is, if there is some positive radius r such that each point $x = (x_1, x_2)$ in the set satisfies

$$\|x\| = \sqrt{x_1^2 + x_2^2} \leq r.$$

Similarly, we can define a bounded set in R^n as follows:

DEFINITION 14.3 A set in R^n is said to be bounded if there is some positive number r such that each point $x = (x_1, x_2, \dots, x_n)$ in the set satisfies

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} \leq r.$$

The following theorem, which is just Theorem 13.1 of Chapter 13 extended to the space R^n , shows that if a linear programming problem has an optimal solution, it can be found among the extreme points of the feasible set.

THEOREM 14.2 If the feasible set of a linear programming problem is nonempty and bounded, then the objective function attains its maximum at an extreme point of the set. If the feasible set is unbounded, then the objective function may or may not attain a maximum value; however, if it attains a maximum value, it does so at an extreme point.

The proof of this theorem follows the same lines as that outlined in Chapter 13 for two variables. Let us apply this theorem to the two problems posed in Examples 14.2 and 14.3.

EXAMPLE 14.2 (REVISITED) As illustrated in Fig. 14.3, the feasible set for this problem is nonempty and bounded. Consequently, by Theorem 14.2 the objective function

$$z = 3x_1 + x_2 - 2x_3$$

attains its maximum at one of the two extreme points of the set. At the extreme point

$$(5/2, 0, 5/2)$$

we have

$$z = 5/2,$$

and at the extreme point

$$(0, 10/3, 20/3)$$

we have

$$z = -10.$$

Thus, the maximum value of the objective function is $z = 5/2$, and an optimal solution to the problem is $x_1 = 5/2$, $x_2 = 0$, and $x_3 = 5/2$.

EXAMPLE 14.3 (REVISITED) The feasible set of Example 14.3 is illustrated in Fig. 14.4. Since it is nonempty and bounded, Theorem 14.2 guarantees that the objective function

$$z = x_1 - 3x_2 + 2x_3$$

attains its maximum value at one of the three extreme points. The following table gives the values of the objective functions at these extreme points:

Extreme point (x_1, x_2, x_3)	Value of $z = x_1 - 3x_2 + 2x_3$
(6, 0, 0)	6
(0, 3, 0)	-9
(0, 0, 4)	8

Thus the maximum value of z is 8 and an optimal solution is $x_1 = 0$, $x_2 = 0$, $x_3 = 4$.

In the above two examples (and in all of the examples in Chapter 13) the extreme points of the feasible sets were found geometrically. But geometric techniques are not possible if the problem has more than three variables. In such cases, we need an algebraic technique for generating the extreme points of the feasible set. In the next section we describe such an algebraic technique.

BASIC FEASIBLE SOLUTIONS

Let us reexamine the linear system

$$Ax = b \quad (14.7)$$

of Problem 14.2, where A is an $m \times n$ matrix. Although it is not essential, we will assume for simplicity that $m \leq n$; i.e. that there are no more constraints than variables. We shall also assume that

the m rows of A are linearly independent. This implies that A contains m linearly independent columns since the row space and column space of any matrix have the same dimension. In Exercise 14.10, we ask the reader to show that the linear system (14.7) can also be written as

$$x_1 a_1 + x_2 a_2 + \cdots + x_n a_n = b \quad (14.8)$$

where a_i ($i = 1, 2, \dots, n$) is the i -th column vector of the matrix A . Thus, solving $Ax = b$ for x is equivalent to solving the vector equation (14.8) for x_1, x_2, \dots, x_n . For convenience, we say that the variable x_i "corresponds" to the column vector a_i . As noted above, there must exist at least one set of m linearly independent vectors among the column vectors a_1, a_2, \dots, a_n . For example, suppose the first m of these vectors, a_1, a_2, \dots, a_m , form such a set. Since these m vectors lie in R^m , and since R^m is m -dimensional, a_1, a_2, \dots, a_m constitute a basis for R^m . Thus the vector b in (14.8) is uniquely expressible as a linear combination of a_1, a_2, \dots, a_m . That is, there is a unique solution of (14.8) for which

$$x_{m+1} = x_{m+2} = \cdots = x_n = 0. \quad (14.9)$$

Similarly, any set of m linearly independent column vectors of A would lead to a solution of (14.8) in which $n - m$ of the variables are zero and the remaining m are uniquely determined. This suggests the following definition:

DEFINITION 14.4 A vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

is called a **basic solution** of the linear system $Ax = b$ if $n - m$ of the variables x_1, x_2, \dots, x_n are zero and the remaining m variables correspond to linearly independent column vectors of A . The $n - m$ zero variables are called the **nonbasic variables**, and the m variables corresponding to the linearly independent column vectors are called the **basic variables** of x .

A linear system of m equations in n variables has as many basic solutions as there are sets of m linearly independent columns among the n columns of the coefficient matrix A .

EXAMPLE 14.4 Find all basic solutions of the linear system

$$\begin{aligned} 2x_1 + x_3 + 4x_4 + 2x_5 &= 20 \\ x_1 + x_2 - x_3 + x_4 + 3x_5 &= 10. \end{aligned}$$

SOLUTION The coefficient matrix of the linear system is

$$A = \begin{bmatrix} 2 & 0 & 1 & 4 & 2 \\ 1 & 1 & -1 & 1 & 3 \end{bmatrix}.$$

As we can see, any two columns of this matrix are linearly independent. Thus to find a basic solution we choose any two of these columns, set the three appropriate nonbasic variables equal to zero and solve the resulting 2×2 linear system for the two basic variables. For example, if we choose the two columns

$$a_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \text{ and } a_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

we set the nonbasic variables

$$x_2, x_4, \text{ and } x_5$$

equal to zero and obtain the system

$$\begin{aligned} 2x_1 + x_3 &= 20 \\ x_1 - x_3 &= 10 \end{aligned}$$

for the basic variables x_1 and x_3 . This system is easily found to have the solution $x_1 = 10$ and $x_3 = 0$. The resulting basic solution to the original problem is

$$x = \begin{bmatrix} 10 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix},$$

where we have underlined the basic variables. As the reader can verify, the ten possible pairs of columns of A lead to the following ten basic solutions:

$$\begin{bmatrix} \underline{10} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \underline{10} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \underline{10} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \underline{10} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \underline{30} \\ 20 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \underline{5} \\ 0 \\ \underline{5} \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ -20 \\ 0 \\ 0 \\ \underline{10} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ -4 \\ \underline{6} \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \underline{8} \\ 0 \\ \underline{6} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \underline{4} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \underline{2} \end{bmatrix} \quad (14.10)$$

where we have underlined the two basic variables in each case.

From (14.10), we see that the first four basic solutions are equal as vectors in the vector space R^5 . Nevertheless, we shall consider them to be distinct basic solutions because they result from different pairs of linearly independent columns of A . This particular circumstance arises because one of the basic variables in each of these four basic solutions is equal to zero. In general, a basic solution is said to be *degenerate* if any of its basic variables is equal to zero. Otherwise, it is said to be *nondegenerate*.

Recall that the feasible set of a standard linear programming problem consists of those vectors which satisfy a linear equation

$$Ax = b$$

and which satisfy the nonnegativity condition.

$$x \geq 0.$$

Adjoining the nonnegativity condition to the concept of a basic solution we are led to the following definition:

DEFINITION 14.5 In a standard linear programming problem a feasible solution which is also a basic solution of the system $Ax = b$ is called a basic feasible solution.

We are now ready to state the following fundamental theorem in the theory of linear programming:

THEOREM 14.3 A vector x is an extreme point of the feasible set of a linear programming problem if and only if it is a basic feasible solution of the problem.

This theorem will yield an algebraic technique for finding extreme points. The proof of this theorem is too long to present here. Instead, we refer the reader to any standard text in linear programming theory, such as S. I. Gass, *Linear Programming*, 4th ed., New York: McGraw-Hill Book Company, 1975.

EXAMPLE 14.5 Find an optimal solution of the linear programming problem:

$$\text{Maximize } z = 2x_1 + 3x_2 - x_3 + 4x_4 + x_5 + x_6$$

subject to

$$2x_1 + x_3 + 4x_4 + 2x_5 = 20$$

$$x_1 + x_2 - x_3 + x_4 + 3x_5 = 10$$

and

$$x_1, x_2, x_3, x_4, x_5 \geq 0.$$

SOLUTION As we ask the reader to show in Exercise 14.9, the feasible set of this problem is bounded. Consequently, from Theorem 14.2 the objective function attains its maximum value at an extreme point. The linear system $Ax = b$ in the constraints is the same system considered in Example 14.4. Equations (14.10) give the ten basic solutions of this linear system. Of these ten, the eight listed in (14.11) are basic feasible solutions since they satisfy the nonnegativity condition. Thus, from Theorem 14.3, the extreme points of the feasible set are also given by Eqs. (14.11). (Notice that these eight basic feasible solutions determine only five distinct extreme points.) Below each of the eight basic feasible solutions we have given the corresponding value of the objective function. From this we see that $z = 70$ is the maximum value of the objective function and an optimal solution is $x_1 = 0, x_2 = 30, x_3 = 20, x_4 = 0, x_5 = 0$.

$$\begin{array}{cccccccc}
 x_1 = \begin{bmatrix} 10 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & x_2 = \begin{bmatrix} 10 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & x_3 = \begin{bmatrix} 10 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & x_4 = \begin{bmatrix} 10 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & x_5 = \begin{bmatrix} 0 \\ 30 \\ 20 \\ 0 \\ 0 \end{bmatrix} & x_6 = \begin{bmatrix} 0 \\ 5 \\ 0 \\ 5 \\ 0 \end{bmatrix} & x_7 = \begin{bmatrix} 0 \\ 0 \\ 8 \\ 0 \\ 6 \end{bmatrix} & x_8 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 4 \\ 2 \end{bmatrix} \\
 z = 20 & z = 20 & z = 20 & z = 20 & z = 70 & z = 35 & z = -2 & z = 18
 \end{array}$$

(14.11)

The technique used in Example 14.5 can, in principle, be used to solve any linear programming problem. However, the number of basic feasible solutions quickly increases as the number of variables increases. For example, a linear programming problem with forty variables and twenty equality constraints could have over 130 billion basic feasible solutions. It would be completely impractical to find all of them, even with the fastest computer. Chapter 15 describes a practical alternative to this technique called the *simplex method*. We finish this chapter with a brief qualitative description of the simplex method to prepare for that chapter.

INTRODUCTION TO THE SIMPLEX METHOD

Let us introduce the following definition:

DEFINITION 14.6 In a linear programming problem, two basic feasible solutions having m basic variables are said to be adjacent if they have $m - 1$ basic variables in common.

Geometrically, extreme points corresponding to adjacent basic feasible solutions are connected by some "edge" of the feasible set. However, we shall not pursue this geometric interpretation.

EXAMPLE 14.5 (REVISITED) Let us find the adjacent basic feasible solutions of the linear programming problem posed in Example 14.5. The eight basic feasible solutions of this problem are given in Eqs. (14.11). In Fig. 14.5 we have drawn a graph of the eight basic feasible solutions. In this graph we have linked adjacent basic feasible solutions with a single line. For example, x_5 and x_7 are adjacent because they have $m - 1 = 2 - 1 = 1$ basic variable in common, namely x_3 . On the other hand, x_4 and x_5 are not adjacent since they do not have $m - 1 = 1$ basic variable in common.

The simplex method is a way of proceeding from one basic feasible solution to an adjacent basic feasible solution in such a way that the value of the objective function never decreases. This usually leads to a basic feasible solution for which the value of the objective function is as large as possible. We say "usually" because there is a slight complication caused by degeneracy which we shall describe below.

In Fig. 14.6(a) we have redrawn Fig. 14.5 and labeled each of the eight basic feasible solutions with the corresponding value of the objective function as given by (14.11). If somehow we generate x_8 as a basic feasible solution, one can show that the simplex

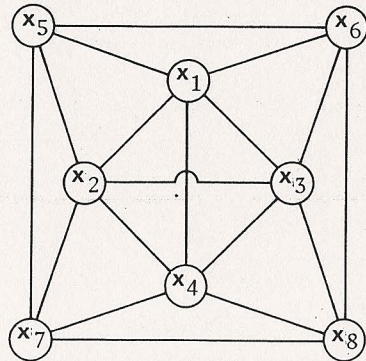


Figure 14.5

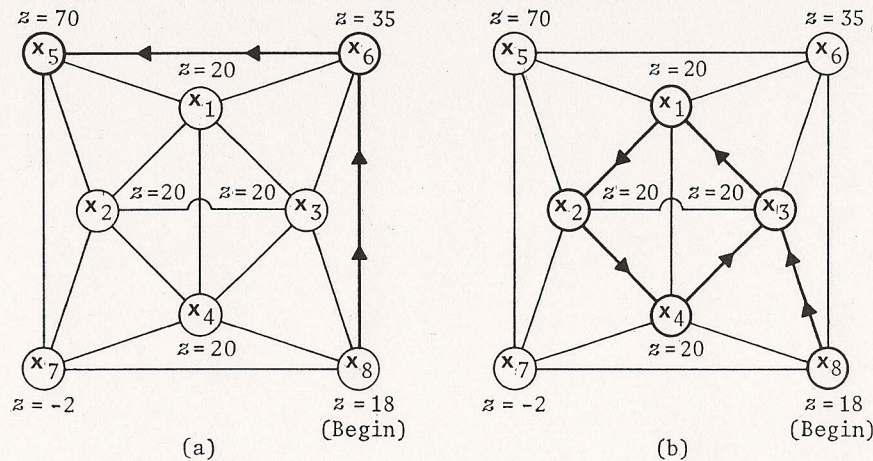


Figure 14.6

method will generate for us the adjacent basic feasible solution x_6 . The value of the objective function is thereby increased from $z = 18$ to $z = 35$. From x_6 , the simplex method will then generate the adjacent basic feasible solution x_5 , where $z = 70$. At all basic feasible solutions adjacent to x_5 , the value of the objective function is less than 70. Thus $z = 70$ is the maximum value of the objective function and x_5 is an optimal solution to the problem.

If any of the basic feasible solutions is degenerate, however, it is possible that the simplex method will not lead to an optimal solution. Since the simplex method only guarantees that the value of the objective function will not decrease, it is possible that the situation described in Figure 14.6(b) might arise. Here, we move around the basic feasible solutions x_3 , x_1 , x_2 , and x_4 indefinitely without ever reaching the optimal basic feasible solution x_5 . This phenomenon is known as *cycling*. Fortunately, it is not a serious problem in realistic linear programming problems. The circumstances which may produce cycling are rarely encountered in practice; and round-off error in a computer tends to destroy degeneracy so that a loop such as in Fig. 14.6(b) is eventually exited. At any rate, there are algorithms available which can eliminate cycling if it is suspected to be encountered in some specific problem.

EXERCISES

14.1 Convert the following linear programming problem to one in standard form:

$$\text{Minimize } z = 2x_1 + 5x_2$$

subject to

$$3x_1 - 6x_2 \leq 2$$

$$x_1 + x_2 \leq 3$$

$$x_1 \geq 6$$

$$x_2 \leq 5$$

and

$$x_1, x_2 \geq 0.$$

14.2 Convert the following linear programming problem to one in standard form:

$$\text{Maximize } z = -3x_1 + x_2 + x_3$$

subject to

$$\begin{aligned} 3x_1 - 5x_2 + x_3 &= 3 \\ -2x_1 - x_2 &\leq 2 \\ x_1 &\geq 5 \\ x_2 &\geq 2 \end{aligned}$$

and

$$x_1, x_2, x_3 \geq 0.$$

14.3 For the following linear programming problem in standard form:

$$\text{Maximize } z = 2x_1 - x_2 + x_3 + 3x_4$$

subject to

$$\begin{aligned} x_1 + x_2 + 2x_3 &= 2 \\ 2x_1 + 4x_3 + x_4 &= 1 \end{aligned}$$

and

$$x_1, x_2, x_3, x_4 \geq 0,$$

- show that the feasible set is bounded,
- find all basic solutions of the linear system $Ax = b$,
- find all basic feasible solutions of the problem,
- find an optimal solution and the maximum value of the objective function,
- draw a graph of the basic feasible solutions, as in Fig. 14.5, in which adjacent basic feasible solutions are linked with a single line.

14.4 For the following linear programming problem in standard form:

$$\text{Maximize } z = 2x_1 - 6x_2 + 3x_3$$

subject to

$$\begin{aligned} x_1 + 5x_2 + 3x_3 &= 2 \\ -x_1 + 2x_2 + 4x_3 &= 3 \end{aligned}$$

and

$$x_1, x_2, x_3 \geq 0,$$

- find all basic solutions of the linear system $Ax = b$,
- show that the problem has no basic feasible solutions. (From this it follows that the feasible set of this problem is empty.)

14.5 For the following linear programming problem:

$$\text{Maximize } z = 3x_1 + 2x_2 - x_3$$

subject to

$$\begin{aligned} 2x_1 + 3x_2 + x_3 &\leq 4 \\ x_1 + 2x_2 + 3x_3 &\leq 5 \end{aligned}$$

and

$$x_1, x_2, x_3 \geq 0,$$

- convert the problem to one in standard form,
- find all basic solutions of the linear system $Ax = b$ in the standard problem,
- find all basic feasible solutions of the standard problem,
- find an optimal solution of the standard problem and the maximum value of the objective function,
- draw a graph of the basic feasible solutions of the standard problem, as in Fig. 14.5, in which adjacent basic feasible solutions are linked with a single line.
- find an optimal solution of the original problem.

14.6 Repeat the instructions in Exercise 14.5 for the linear programming problem:

$$\text{Minimize } z = 2x_1 - 3x_2 + x_3$$

subject to

$$\begin{aligned} x_1 - 2x_2 + 3x_3 &\leq 5 \\ 2x_1 + x_2 - 2x_3 &= 2 \end{aligned}$$

and

$$x_1, x_2, x_3 \geq 0.$$

14.7 Prove Theorem 14.1 as follows:

- Show that if $Ax_1 = b$ and $Ax_2 = b$, then $Ax = b$ if $x = tx_1 + (1-t)x_2$ for any t in the interval $[0, 1]$.
- Show that if $x_1 \geq 0$ and $x_2 \geq 0$, then $x \geq 0$ if $x = tx_1 + (1-t)x_2$ for any t in the interval $[0, 1]$.

14.8 Show that the feasible set of the linear programming problem in Example 14.2 is the set illustrated in Fig. 14.3.

14.9 Show that the feasible set of the linear programming problem in Example 14.5 is bounded. Proceed as follows:

(a) From the constraint

$$2x_1 + x_3 + 4x_4 + 2x_5 = 20$$

and the nonnegativity conditions $x_i \geq 0$ ($i = 1, 2, 3, 4, 5$) conclude that $x_i \leq 20$ for $i = 1, 3, 4, 5$

(b) Add the two constraints together and conclude that $x_2 \leq 30$.

(c) From (a) and (b), conclude that $\|x\| \leq r$ for some positive number r .

14.10 Show that the linear system (14.7) can be written in the vector form (14.8).

Linear Programming 3: The Simplex Method

The simplex method is presented and applied to the solution of a certain class of linear programming problems.

PREREQUISITES: Gaussian elimination
Chapter 14: Linear Programming II

INTRODUCTION

At the end of Chapter 14, we described the simplex method as a procedure for moving from one basic feasible solution of a linear programming problem to an adjacent basic feasible solution in such a way that the value of the objective function never decreases. In this chapter we shall describe the algebraic details of this procedure.

The general linear programming problem in n variables was stated as Problem 14.1 on page 182. However, in order to simplify the presentation of the simplex method, we shall restrict ourselves to linear programming problems having the following special form:

PROBLEM 15.1 Find values of x_1, x_2, \dots, x_n which maximize

$$z = c_1x_1 + c_2x_2 + \dots + c_nx_n$$

subject to

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \leq b_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \leq b_2$$

⋮

$$a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n \leq b_m$$

and,

$$x_i \geq 0, \quad \text{for } i = 1, 2, \dots, n,$$

where

$$b_j \geq 0, \quad \text{for } j = 1, 2, \dots, m.$$

In Problem 15.1, the condition that each of the m constraints be a \leq inequality is not restrictive since it can easily be shown that any linear programming problem can always be written with all \leq constraints. It is the condition $b_j \geq 0$ for $j = 1, 2, \dots, m$ that is the real restriction. Nevertheless, a large class of practical problems are of this form, and the procedures developed in this chapter for Problem 15.1 are used in the application of the simplex method to the general linear programming problem.

To convert Problem 15.1 to one in standard form (see Problem 14.2 on page 182) we introduce m slack variables $x_{n+1}, x_{n+2}, \dots, x_{n+m}$, one for each of the m constraints, to obtain

PROBLEM 15.2 Find values of x_1, x_2, \dots, x_{n+m} which maximize

$$z = c_1x_1 + c_2x_2 + \dots + c_nx_n + 0x_{n+1} + \dots + 0x_{n+m}$$

subject to

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n + x_{n+1} = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n + x_{n+2} = b_2$$

⋮

$$a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n + x_{n+m} = b_m$$

and

$$x_i \geq 0, \quad \text{for } i = 1, 2, \dots, n+m.$$

If we can find an optimal solution to Problem 15.2, then the values of the variables x_1, x_2, \dots, x_n will provide an optimal solution to problem 15.1.

THE SIMPLEX TABLEAU

In order to more clearly describe the steps in the simplex method, let us examine the following specific problem of the type we are considering in this chapter:

PROBLEM 15.3 Find values of x_1, x_2 , and x_3 which maximize

$$z = 3x_1 - x_2 + 4x_3$$

subject to

$$2x_1 - x_2 + 3x_3 \leq 5$$

$$x_1 + 4x_2 - 2x_3 \leq 1$$

$$3x_1 + 6x_3 \leq 4$$

and

$$x_1, x_2, x_3 \geq 0.$$

To convert Problem 15.3 to one in standard form, we add slack variables x_4, x_5, x_6 to obtain

PROBLEM 15.4 Find values of $x_1, x_2, x_3, x_4, x_5,$ and x_6 which maximize

$$z = 3x_1 - x_2 + 4x_3 + 0x_4 + 0x_5 + 0x_6$$

subject to

$$2x_1 - x_2 + 3x_3 + x_4 = 5$$

$$x_1 + 4x_2 - 2x_3 + x_5 = 1$$

$$3x_1 + 6x_3 + x_6 = 4$$

and

$$x_1, x_2, x_3, x_4, x_5, x_6 \geq 0.$$

For our purposes, it will be convenient to rephrase Problem 15.4 in the following equivalent form:

PROBLEM 15.5 Find values of $x_1, x_2, x_3, x_4, x_5, x_6,$ and z which satisfy

$$\begin{aligned} 2x_1 - x_2 + 3x_3 + x_4 &= 5 \\ x_1 + 4x_2 - 2x_3 + x_5 &= 1 \\ 3x_1 + 6x_3 + x_6 &= 4 \\ -3x_1 + x_2 - 4x_3 + z &= 0 \end{aligned} \quad (15.1)$$

and such that $x_1, x_2, x_3, x_4, x_5, x_6$ are nonnegative and z is as large as possible.

In this formulation of the problem, z is treated as a variable on a par with x_1 through x_6 , and the equation defining z in terms of the x_i is treated as an additional constraint. Thus our problem is to find a solution of the linear system of four equations in seven unknowns given by (15.1) in which one of the variables, z , is as large as possible and the other six variables are nonnegative.

The usual procedure for solving a linear system of equations is to construct the augmented matrix of the system and apply Gaussian elimination or Gauss-Jordan elimination to it to put it in row-echelon form (or reduced row-echelon form). The row-echelon

form then determines the augmented matrix of an equivalent linear system which is easily solved. The simplex method proceeds along similar lines. Namely, the augmented matrix of the linear system is constructed and a variation of Gaussian elimination, called *pivotal elimination*, is applied to obtain augmented matrices in which basic feasible solutions to the linear programming problem can be determined by inspection. Let us return to Problem 15.5 to see how this is done.

The augmented matrix of the linear system (15.1) is

$$\left[\begin{array}{ccccccc|c} 2 & -1 & 3 & 1 & 0 & 0 & 0 & 5 \\ 1 & 4 & -2 & 0 & 1 & 0 & 0 & 1 \\ 3 & 0 & 6 & 0 & 0 & 1 & 0 & 4 \\ -3 & 1 & -4 & 0 & 0 & 0 & 1 & 0 \end{array} \right] \quad (15.2)$$

From the way the 1's and 0's are distributed in the 4th, 5th, 6th, and 7th columns of this matrix, one particular solution of (15.1) can be seen by inspection; namely,

$$x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 5, x_5 = 1, x_6 = 4, z = 0. \quad (15.3)$$

In terms of the corresponding linear programming problem, this is

$$x' = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 5 \\ 1 \\ 4 \end{bmatrix} \quad \text{and} \quad z = 0. \quad (15.4)$$

It is easily seen that x' is a basic feasible solution of linear programming Problem 15.4, according to Definition 14.5 on page 193. The three variables x_4, x_5, x_6 are the basic variables, and their corresponding values are found as the first three entries of the last column of (15.2). The value of z is the last entry of this column. That x' is a basic feasible solution is of crucial importance, since it is among the basic feasible solutions that we can hope to find an optimal solution. To see how we may go about finding another basic feasible solution, we rewrite (15.2) with some additional labeling:

Tableau 15.1

x_1	x_2	x_3	x_4	x_5	x_6	z	
2	-1	3	1	0	0	0	5 = x_4
1	4	-2	0	1	0	0	1 = x_5
3	0	6	0	0	1	0	4 = x_6
-3	1	-4	0	0	0	1	0 = z

Each column has been labeled with its corresponding variable from the linear system (15.1). On the right, we have labeled the entries of the last column with the corresponding variables whose values they determine in the solution given by (15.3). We have also drawn a vertical and horizontal line within the matrix in order to highlight certain entries which will be useful to us later on.

In the field of linear programming, the augmented matrix is referred to as a *tableau*. In particular, Tableau 15.1 above is called the *initial tableau* of the problem. We shall call the last row of the tableau the *objective row*, since it arises from the objective function of the original problem.

In Tableau 15.1 we have also shaded four particular columns. It can be seen that these four columns are identical to the columns of the 4×4 identity matrix. Indeed, it was exactly this fact that permitted us to find the solution given in (15.3) so easily. We had only to set those variables not corresponding to these four columns equal to zero, and then the values of the variables corresponding to the four columns were found in the last column of the augmented matrix. This suggests a way of proceeding to a new solution of the linear system. We apply appropriate elementary row operations to Tableau 15.1 to arrive at a new tableau which again contains the four columns of the 4×4 identity matrix, but this time in different position. To see how to do this, consider the following tableau, which is just Tableau 15.1 with one of its entries shaded. (We postpone for the moment a discussion of why this particular entry was chosen.)

Tableau 15.2

x_1	x_2	x_3	x_4	x_5	x_6	z	
2	-1	3	1	0	0	0	5 = x_4
1	4	-2	0	1	0	0	1 = x_5
3	0	6	0	0	1	0	4 = x_6
-3	1	-4	0	0	0	1	0 = z

Our objective will be to use elementary row operations to replace the shaded entry by a "1" and obtain zeros everywhere else in that column. To do this we first divide the third row by six to obtain:

Tableau 15.3

x_1	x_2	x_3	x_4	x_5	x_6	z	
2	-1	3	1	0	0	0	5
1	4	-2	0	1	0	0	1
1/2	0	1	0	0	1/6	0	2/3
-3	1	-4	0	0	0	1	0

Next we perform the following three elementary row operations:

1. Add -3 times the 3rd row to the 1st row.
2. Add 2 times the 3rd row to the 2nd row.
3. Add 4 times the 3rd row to the 4th row.

The result is the following tableau:

Tableau 15.4

x_1	x_2	x_3	x_4	x_5	x_6	z	
1/2	-1	0	1	0	-1/2	0	3 = x_4
2	4	0	0	1	1/3	0	7/3 = x_5
1/2	0	1	0	0	1/6	0	2/3 = x_3
-1	0	0	0	0	2/3	1	8/3 = z

It can be seen that Tableau 15.4 contains within it the four columns of the 4×4 identity matrix, though not in their usual order. Consequently, if we set the variables not associated with these columns equal to zero, we obtain the following solution to linear system (15.1):

$$x_1 = 0, x_2 = 0, x_3 = 2/3, x_4 = 3, x_5 = 7/3, x_6 = 0, z = 8/3. \quad (15.5)$$

As before, we have labeled the entries in the last column with the variables whose values they determine in this solution.

Solution (15.5) specifies the following vector solution of Problem 15.4:

$$x'' = \begin{bmatrix} 0 \\ 0 \\ 2/3 \\ 3 \\ 7/3 \\ 0 \end{bmatrix} \quad \text{and } z = 8/3. \quad (15.6)$$

We see that x'' is a basic feasible solution with basic variables x_3, x_4, x_5 . We also see that x' as given in (15.4) and x'' are *adjacent basic feasible solutions* since they have $m-1=2$ basic variables in common; namely, x_4 and x_5 . That is, the elementary row operations we performed on Tableau 15.1 to obtain Tableau 15.4 took us from the basic feasible solution x' to the adjacent basic feasible solution x'' . In addition, the value of the objective function increased from $z = 0$ to $z = 8/3$.

The shaded entry in Tableau 15.2 is called the *pivot entry* of that particular tableau. The column in which it lies is called the *pivot column*, and the row in which it lies is called the *pivot row*. After the elementary row operations were applied to obtain Tableau 15.4, the pivot column contained all zeros, except for a one in the pivot entry position. Thus the pivot column was converted to one of the columns of the 4×4 identity matrix, and the variable x_3 corresponding to the pivot column was converted from a nonbasic variable to a basic variable. At the same time, the variable x_6 which labeled the pivot row, was converted from a basic variable to a nonbasic variable. That is, x_3 replaced x_6 as a basic variable in going from Tableau 15.1 to Tableau 15.4. For this reason, x_3 is called the *entering variable* and x_6 is called the *departing variable* of Tableau 15.1.

Let us see if we can increase the value of the objective function above the value $z = 8/3$ attained in Tableau 15.4. Below we have rewritten Tableau 15.4 with a pivot entry shaded. Again, we postpone until the next section a discussion of how this pivot entry was chosen.

Tableau 15.5

	x_1	x_2	x_3	x_4	x_5	x_6	z	
Depart	1/2	-1	0	1	0	-1/2	0	3 = x_4
x_5 ←	2	4	0	0	1	1/3	0	7/3 = x_5
	1/2	0	1	0	0	1/6	0	2/3 = x_3
	-1	0	0	0	0	2/3	1	8/3 = z
↑								Enter x_1

We apply the elementary row operations to the first column to convert it to a column with a "1" in the pivot position and zeros everywhere else. Thus, x_1 will be the entering variable and x_5 will be the departing variable, as the arrows in the tableau indicate. The reader can easily verify that the elementary row operations necessary will produce the next tableau:

Tableau 15.6

x_1	x_2	x_3	x_4	x_5	x_6	z	
0	-2	0	1	-1/4	-7/12	0	29/12 = x_4
1	2	0	0	1/2	1/6	0	7/6 = x_1
0	-1	1	0	-1/4	1/12	0	1/12 = x_3
0	2	0	0	1/2	5/6	1	23/6 = z

Setting those variables not corresponding to the shaded columns equal to zero yields the solution

$$x_1 = 7/6, x_2 = 0, x_3 = 1/12, x_4 = 29/12, x_5 = 0, x_6 = 0, z = 23/6. \quad (15.7)$$

For the linear programming problem, we then have the following basic feasible solution and objective function value:

$$x''' = \begin{bmatrix} 7/6 \\ 0 \\ 1/12 \\ 29/12 \\ 0 \\ 0 \end{bmatrix} \quad \text{and } z = 23/6. \quad (15.8)$$

As we shall show in the next section, the value $z = 23/6$ is the largest value the objective function can assume over the feasible set. Thus we have reached an optimal solution. For the solution to Problem 15.3, we discard the slack variables x_4, x_5, x_6 and write

$$x_1 = 7/6, x_2 = 0, x_3 = 1/12 \quad (15.9)$$

as the optimal solution, with the corresponding maximum $z = 23/6$ for the objective function.

This example illustrates the kinds of calculations required to implement the simplex method. In the next section we shall discuss how to choose the pivot entry and how to determine if an optimal solution has been reached.

STEPS IN THE SIMPLEX METHOD

In this section we shall outline the steps in the simplex method and give an example. In the next section we shall discuss their mathematical justification. The simplex method consists of the following five steps:

- STEP 1 *Construct the initial tableau.*
 STEP 2 *Test for optimality. If the tableau yields an optimal solution, then stop; otherwise, continue to Step 3.*
 STEP 3 *Determine the pivot column.*
 STEP 4 *Determine the pivot row.*
 STEP 5 *Apply the elementary row operations to obtain all zeros in the pivot column, except for a "one" in the pivot row. Return to Step 2.*

The details for Steps 2, 3, and 4 are as follows:

Test for Optimality. If all of the entries in the objective row are nonnegative (ignoring the rightmost entry) the tableau determines an optimal solution.

Determination of Pivot Column. Choose the pivot column so that it contains the most negative entry of the objective row (ignoring the rightmost entry).

Determination of Pivot Row. Ignoring the objective row, divide each positive entry of the pivot column into the last entry in its row. Choose the pivot row to be one which yields the smallest such ratio.

The reader should return to the tableaus constructed in connection with Problem 15.5 to verify that the pivot entries in each tableau were selected according to the above rules, and also to verify that Tableau 15.6 determines an optimal solution to the problem.

Let us apply the simplex method as described above to the following example:

EXAMPLE 15.1 Find values of x_1, x_2, x_3 which maximize

$$z = 3x_1 + 4x_2 + 2x_3$$

subject to

$$3x_1 + 2x_2 + 4x_3 \leq 15$$

$$x_1 + 2x_2 + 3x_3 \leq 7$$

$$2x_1 + x_2 + x_3 \leq 6$$

and

$$x_1, x_2, x_3 \geq 0.$$

SOLUTION In standard form, this problem is

$$\text{Maximize } z = 3x_1 + 4x_2 + 2x_3 + 0x_4 + 0x_5 + 0x_6$$

subject to

$$3x_1 + 2x_2 + 4x_3 + x_4 = 15$$

$$x_1 + 2x_2 + 3x_3 + x_5 = 7$$

$$2x_1 + x_2 + x_3 + x_6 = 6$$

and

$$x_1, x_2, x_3, x_4, x_5, x_6 \geq 0.$$

The initial tableau for the problem is then

Tableau 15.7

x_1	x_2	x_3	x_4	x_5	x_6	z	
3	2	4	1	0	0	0	15 = x_4
1	2	3	0	1	0	0	7 = x_5
2	1	1	0	0	1	0	6 = x_6
-3	-4	-2	0	0	0	1	0 = z

The objective row contains negative entries, so that the initial tableau does not determine an optimal solution. The most negative entry, -4, lies in the second column, so that the second column will be the pivot column. To determine the pivot row, we evaluate the following ratios:

1st row: $15/2 = 7\frac{1}{2}$. 2nd row: $7/2 = 3\frac{1}{2}$. 3rd row: $6/1 = 6$.

The 2nd row yields the smallest ratio so that it will be the pivot row. So far we have the following:

Tableau 15.8

	x_1	x_2	x_3	x_4	x_5	x_6	z	
Depart x_5 ←	3	2	4	1	0	0	15	$= x_4$
	1	2	3	0	1	0	7	$= x_5$
	2	1	1	0	0	1	6	$= x_6$
	-3	-4	-2	0	0	0	0	$= z$

↑
Enter x_2

We now apply the following elementary row operations to Tableau 15.8:

1. Divide the 2nd row by 2.
2. Add -2 times the 2nd row to the 1st row.
3. Add -1 times the 2nd row to the 3rd row.
4. Add 4 times the 2nd row to the 4th row.

The resulting tableau is the following:

Tableau 15.9

	x_1	x_2	x_3	x_4	x_5	x_6	z	
	2	0	1	1	-1	0	8	$= x_4$
	1/2	1	3/2	0	1/2	0	7/2	$= x_2$
	3/2	0	-1/2	0	-1/2	1	5/2	$= x_6$
	-1	0	4	0	2	0	14	$= z$

The objective row still contains a negative entry, so that we have not yet reached an optimal solution. The new pivot column is the first since it contains the only negative entry of the objective row. To determine the pivot row, we evaluate the following ratios:

1st row: $8/2 = 4$. 2nd row: $(7/2)/(1/2) = 7$. 3rd row: $(5/2)/(3/2) = 1\frac{2}{3}$.

The third row yields the smallest ratio, so that it is the new pivot row. So far, we have

Tableau 15.10

	x_1	x_2	x_3	x_4	x_5	x_6	z	
	2	0	1	1	-1	0	8	$= x_4$
	1/2	1	3/2	0	1/2	0	7/2	$= x_2$
Depart x_6 ←	3/2	0	-1/2	0	-1/2	1	5/2	$= x_6$
	-1	0	4	0	2	0	14	$= z$

↑
Enter x_1

We now apply the following elementary row operations to Tableau 15.10:

1. Divide the 3rd row by 3/2.
2. Add -2 times the 3rd row to the 1st row.
3. Add $-\frac{1}{2}$ times the 3rd row to the 2nd row.
4. Add 1 times the 3rd row to the 4th row.

The resulting tableau is

Tableau 15.11

	x_1	x_2	x_3	x_4	x_5	x_6	z	
	0	0	5/3	1	-1/3	-4/3	14/3	$= x_4$
	0	1	5/3	0	2/3	-1/3	8/3	$= x_2$
	1	0	-1/3	0	-1/3	2/3	5/3	$= x_1$
	0	0	11/3	0	5/3	2/3	47/3	$= z$

The objective row of this tableau does not contain any negative entries and so this tableau determines an optimal solution. The basic variables in the optimal basic feasible solution are x_1 , x_2 , and x_4 as the righthand labeling indicates. The optimal solution is

$$x_1 = 5/3, x_2 = 8/3, x_3 = 0, x_4 = 14/3, x_5 = 0, x_6 = 0, z = 47/3.$$

For the original problem posed in this example, we discard the slack variables x_4, x_5, x_6 and simply write

$$x_1 = 5/3, \quad x_2 = 8/3, \quad x_3 = 0, \quad z = 47/3.$$

We conclude this section with some remarks concerning complications which may arise in the use of the simplex method as we have described it:

1. In Step 3 it is possible that there is a tie for the most negative entry in the objective row. In that case, any one of them may be chosen, and no complications arise.
2. In Step 4 it is possible that there is more than one row with the same smallest ratio. In that case, any one of them may be used to determine the pivot row and no complications arise in the calculations. However, if such a tie arises, it can be shown that the basic feasible solution determined by the next tableau will be degenerate (i.e., will have a basic variable whose value is zero). As discussed in the last chapter, it is degeneracy which may bring about cycling. But as we also mentioned, it is more a theoretical problem than a practical problem.
3. In Step 4 it is possible that no entry in the pivot column is positive, in which case our technique for evaluating the pivot row is meaningless. It can be shown that if this situation arises, the problem has an *unbounded solution*.

JUSTIFICATION OF THE STEPS IN THE SIMPLEX METHOD

Let us return to the linear programming problem in $n+m$ variables posed in Problem 15.2. Suppose at some point in our calculations we have arrived at Tableau 15.12. (In Exercise 15.11, we ask the reader to show that in any tableau the column labeled "z" always has the form indicated.) Thus the current basic variables are $x_{B1}, x_{B2}, \dots, x_{Bm}$ with corresponding values d_1, d_2, \dots, d_m , and the current value of the objective variable is w . Let us see if the entry y_{rs} would make a suitable pivot entry. The entering variable would be x_s and the departing variable would be x_{Br} . If $y_{rs} \neq 0$, the elementary row operations in Step 5 of the simplex method will produce a tableau having the form of Tableau 15.13.

Tableau 15.12

	x_1	x_2	\dots	x_s	\dots	x_{n+m}	z		
	y_{11}	y_{12}	\dots	y_{1s}	\dots	$y_{1,n+m}$	0	d_1	$= x_{B1}$
	y_{21}	y_{22}	\dots	y_{2s}	\dots	$y_{2,n+m}$	0	d_2	$= x_{B2}$
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
Depart	y_{r1}	y_{r2}	\dots	y_{rs}	\dots	$y_{r,n+m}$	0	d_r	$= x_{Br}$
x_{Br} ←	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	y_{m1}	y_{m2}	\dots	y_{ms}	\dots	$y_{m,n+m}$	0	d_m	$= x_{Bm}$
	c_1	c_2	\dots	c_s	\dots	c_{n+m}	1	w	$= z$

↑
Enter x_s

Tableau 15.13

	x_1	x_2	\dots	x_s	\dots	x_{n+m}	z		
	y_{11}^*	y_{12}^*	\dots	0	\dots	$y_{1,n+m}^*$	0	$d_1 - y_{1s} d_r / y_{rs}$	$= x_{B1}$
	y_{21}^*	y_{22}^*	\dots	0	\dots	$y_{2,n+m}^*$	0	$d_2 - y_{2s} d_r / y_{rs}$	$= x_{B2}$
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	y_{r1}^*	y_{r2}^*	\dots	1	\dots	$y_{r,n+m}^*$	0	d_r / y_{rs}	$= x_s$
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	y_{m1}^*	y_{m2}^*	\dots	0	\dots	$y_{m,n+m}^*$	0	$d_m - y_{ms} d_r / y_{rs}$	$= x_{Bm}$
	c_1^*	c_2^*	\dots	0	\dots	c_{n+m}^*	1	$w - c_s d_r / y_{rs}$	$= z$

The pivot column now contains all zeros except for the "1" in the previous pivot entry position. All of the other entries in the tableau have new values, which we indicate with asterisks, except

in the last column where we have explicitly written the new values in terms of the entries in Tableau 15.12. First, let us examine the value of the entering variable x_s :

$$x_s = d_r / y_{rs} . \quad (15.10)$$

Since $d_r \geq 0$, we see that we must have $y_{rs} > 0$ in order to satisfy the constraint $x_s \geq 0$. Let us list this fact as

OBSERVATION 15.1 *The pivot entry must be positive in order that the new tableau determine a feasible solution.*

The remaining basic variables have values given by

$$x_{Bi} = d_i - y_{is} d_r / y_{rs} \quad \text{for } i = 1, 2, \dots, m; i \neq s. \quad (15.11)$$

We must have $x_{Bi} \geq 0$ for the new solution to be feasible. Now if for some i we have $y_{is} \leq 0$, then (15.11) states that for that i , $x_{Bi} \geq 0$ since $d_i \geq 0$, $d_r \geq 0$, and $y_{rs} > 0$. On the other hand, if for some i we have $y_{is} > 0$, then (15.11) requires that

$$d_i - y_{is} d_r / y_{rs} \geq 0 \quad (15.12)$$

in order that $x_{Bi} \geq 0$ for that i . Equation (15.12) can also be written as

$$\frac{d_r}{y_{rs}} \leq \frac{d_i}{y_{is}} . \quad (15.13)$$

In other words, Eq. (15.13) must be satisfied for all those i for which $y_{is} > 0$ in order that the new tableau determine a feasible solution. We state this as

OBSERVATION 15.2 *In order that the new tableau determine a feasible solution, the following must be true: The ratio of the element in the rightmost column of the pivot row to the pivot entry must be the smallest of the corresponding ratios in all of the other rows which contain positive entries in the pivot column (ignoring the objective row).*

Next, let us examine the new value w^* of the objective function. From Tableau 15.13, we see that

$$w^* = w - c_s d_r / y_{rs} . \quad (15.14)$$

Ideally, we would want the increase in the objective function

$$w^* - w = -c_s d_r / y_{rs} \quad (15.15)$$

to be as large as possible. But this would require that we compute the quantities

$$-c_s d_r / y_{rs}$$

for all possible values of r and s to find the largest one. Usually this is not done because of the large number of calculations this would require. Instead, the entering variable x_s is chosen so that c_s is as negative as possible. Since $d_r \geq 0$ and $y_{rs} > 0$, Eq. (15.15) then guarantees that

$$w^* - w \geq 0 ;$$

i.e., that the objective function does not decrease. As discussed in the last chapter, this always eventually leads to an optimal solution, except for the very remote possibility of cycling. Let us summarize this as follows:

OBSERVATION 15.3 *Choose the pivot column so that it contains the most negative entry of the objective row (ignoring the entry in the rightmost column).*

Equation (15.15) also tells us the following: If all of the c 's are nonnegative then the value of the objective function cannot increase, regardless of the choice of pivot entry. In this case, we must have already attained the maximum value of the objective function. Thus, we have

OBSERVATION 15.4 *If all of the entries in the objective row, except for the rightmost entry, are nonnegative, the tableau determines an optimal solution.*

The above four observations justify the steps in the simplex method.

Readers interested in pursuing linear programming in more detail are referred to the following texts:

S. I. Gass, *Linear Programming*, 4th ed. New York: McGraw-Hill Book Company, 1975.

L. Cooper and D. Steinberg, *Methods and Applications of Linear Programming*, Philadelphia: W. B. Saunders Company, 1974.

EXERCISES

In Exercises 15.1 to 15.6 solve the given linear programming problem by the simplex method.

15.1 Maximize $z = 3x_1 + 4x_2$

subject to

$$2x_1 + 3x_2 \leq 7$$

$$5x_1 + 2x_2 \leq 3$$

and

$$x_1, x_2 \geq 0.$$

15.2 Maximize $z = 2x_1 + x_2$

subject to

$$3x_1 + 2x_2 \leq 4$$

$$3x_1 + x_2 \leq 3$$

$$2x_1 \leq 3$$

and

$$x_1, x_2 \geq 0.$$

15.3 Maximize $z = 3x_1 - 2x_2 + 6x_3$

subject to

$$2x_1 - 5x_2 + x_3 \leq 2$$

$$x_1 + x_2 + x_3 \leq 5$$

and

$$x_1, x_2, x_3 \geq 0.$$

15.4 Maximize $z = 2x_1 + x_2 - x_3$

subject to

$$2x_1 - 3x_2 + x_3 \leq 2$$

$$x_1 + 5x_2 - 2x_3 \leq 4$$

$$2x_1 - 4x_2 - x_3 \leq 3$$

and

$$x_1, x_2, x_3 \geq 0.$$

15.5 Maximize $z = 3x_1 - 2x_2 - x_3 + x_4$

subject to

$$2x_1 - 3x_2 + x_3 - x_4 \leq 6$$

$$x_1 + 2x_2 - x_3 + 2x_4 \leq 4$$

and

$$x_1, x_2, x_3, x_4 \geq 0$$

15.6 Maximize $z = x_1 - 2x_2 + 3x_3 + x_4$

subject to

$$x_1 - 2x_2 + x_3 + 3x_4 \leq 8$$

$$2x_1 + 3x_2 - x_3 + 2x_4 \leq 5$$

$$x_1 + x_2 - 3x_3 + 4x_4 \leq 6$$

and

$$x_1, x_2, x_3, x_4 \geq 0$$

15.7 Solve Example 13.1 on page 163 by the simplex method.

15.8 Solve Exercise 13.1 on page 177 by the simplex method.

15.9 Solve Exercise 13.6 on page 179 by the simplex method.

15.10 Solve Exercise 13.7 on page 179 by the simplex method.

15.11 Show that in any tableau the column labeled "z" always has the form indicated in Tableau 15.12.

Answers to Exercises

Chapter 1, page 8

1.1 (a) $y = 3x - 4$

(b) $y = -2x + 1$

1.2 (a) $x^2 + y^2 - 4x - 6y + 4 = 0$ or $(x - 2)^2 + (y - 3)^2 = 9$

(b) $x^2 + y^2 + 2x - 4y - 20 = 0$ or $(x + 1)^2 + (y - 2)^2 = 25$

1.3 $4x^2 + 8xy + 4y^2 - 8x + 4y = 0$ (a parabola)

1.4 (a) $x + 2y + z = 0$

(b) $-x + y - 2z + 1 = 0$

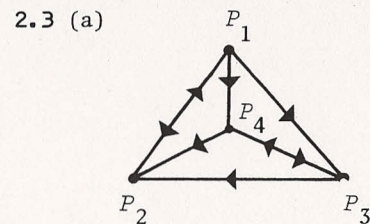
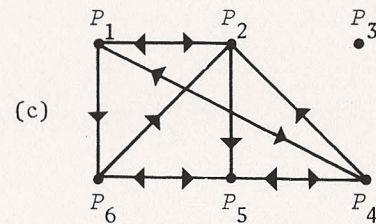
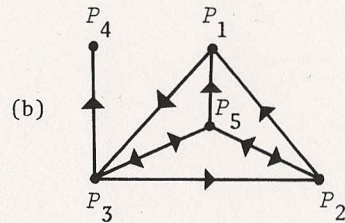
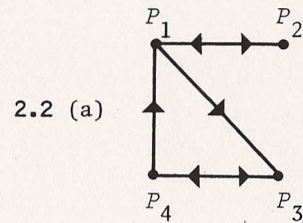
1.5 (a) $x^2 + y^2 + z^2 - 2x - 4y - 2z = -2$ or $(x - 1)^2 + (y - 2)^2 + (z - 1)^2 = 4$

(b) $x^2 + y^2 + z^2 - 6x - 2y - 4z = -5$ or $(x - 3)^2 + (y - 1)^2 + (z - 2)^2 = 9$

1.9
$$\begin{vmatrix} y & x^2 & x & 1 \\ y_1 & x_1^2 & x_1 & 1 \\ y_2 & x_2^2 & x_2 & 1 \\ y_3 & x_3^2 & x_3 & 1 \end{vmatrix} = 0$$

Chapter 2, page 22

2.1 (a) $\begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$



- (b) 1-step: $P_1 \rightarrow P_2$
 2-step: $P_1 \rightarrow P_4 \rightarrow P_2$
 $P_1 \rightarrow P_3 \rightarrow P_2$
 3-step: $P_1 \rightarrow P_2 \rightarrow P_1 \rightarrow P_2$
 $P_1 \rightarrow P_3 \rightarrow P_4 \rightarrow P_2$
 $P_1 \rightarrow P_4 \rightarrow P_3 \rightarrow P_2$

- (c) 1-step: $P_1 \rightarrow P_4$
 2-step: $P_1 \rightarrow P_3 \rightarrow P_4$
 3-step: $P_1 \rightarrow P_2 \rightarrow P_1 \rightarrow P_4$
 $P_1 \rightarrow P_4 \rightarrow P_3 \rightarrow P_4$

- 2.4 (a) $\{P_1, P_2, P_3\}$ (b) $\{P_3, P_4, P_5\}$ (c) $\{P_2, P_4, P_6, P_8\}$ and $\{P_4, P_5, P_6\}$

- 2.5 (a) None (b) $\{P_3, P_4, P_6\}$

2.6 $\begin{bmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$ Power of $P_1 = 5$
 Power of $P_2 = 3$
 Power of $P_3 = 4$
 Power of $P_4 = 2$

- 2.7 First, A; second, B and E (tie); fourth, C; fifth, D.

Chapter 3, page 37

- 3.1 (a) $-5/8$

(b) $\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}^t$

- 3.2 Let $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ for example.

3.3 (a) $p^* = \begin{bmatrix} 0 & 1 \end{bmatrix}$, $q^* = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $v = 3$

(b) $p^* = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$, $q^* = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $v = 2$

(c) $p^* = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$, $q^* = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, $v = 2$

(d) $p^* = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$, $q^* = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $v = -2$

3.4 (a) $p^* = \begin{bmatrix} 5/8 & 3/8 \end{bmatrix}$, $q^* = \begin{bmatrix} 1/8 \\ 7/8 \end{bmatrix}$, $v = 7/4$

(b) $p^* = \begin{bmatrix} 2/3 & 1/3 \end{bmatrix}$, $q^* = \begin{bmatrix} 1/6 \\ 5/6 \end{bmatrix}$, $v = 70/3$

(c) $p^* = \begin{bmatrix} 1 & 0 \end{bmatrix}$, $q^* = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $v = 3$

$$(d) p^* = \begin{bmatrix} 3/5 & 2/5 \end{bmatrix}, \quad q^* = \begin{bmatrix} 3/5 \\ 2/5 \end{bmatrix}, \quad v = 19/5$$

$$(e) p^* = \begin{bmatrix} 3/13 & 10/13 \end{bmatrix}, \quad q^* = \begin{bmatrix} 1/13 \\ 12/13 \end{bmatrix}, \quad v = -29/13$$

$$3.5 \quad p^* = \begin{bmatrix} 7/20 & 13/20 \end{bmatrix}, \quad q^* = \begin{bmatrix} 9/20 \\ 11/20 \end{bmatrix}, \quad v = -3/20$$

Chapter 4, page 51

$$4.1 (a) \quad x^{(1)} = \begin{bmatrix} .4 \\ .6 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} .46 \\ .54 \end{bmatrix}, \quad x^{(3)} = \begin{bmatrix} .454 \\ .546 \end{bmatrix},$$

$$x^{(4)} = \begin{bmatrix} .4546 \\ .5454 \end{bmatrix}, \quad x^{(5)} = \begin{bmatrix} .45454 \\ .54546 \end{bmatrix}$$

(b) P is regular since all entries of P are positive;

$$q = \begin{bmatrix} 5/11 \\ 6/11 \end{bmatrix}$$

$$4.2 (a) \quad x^{(1)} = \begin{bmatrix} .7 \\ .2 \\ .1 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} .23 \\ .52 \\ .25 \end{bmatrix}, \quad x^{(3)} = \begin{bmatrix} .273 \\ .396 \\ .331 \end{bmatrix}$$

(b) P is regular since all entries of P are positive;

$$q = \begin{bmatrix} 22/72 \\ 29/72 \\ 21/72 \end{bmatrix}$$

$$4.3 (a) \quad \begin{bmatrix} 9/17 \\ 8/17 \end{bmatrix} \quad (b) \quad \begin{bmatrix} 26/45 \\ 19/45 \end{bmatrix} \quad (c) \quad \begin{bmatrix} 3/19 \\ 4/19 \\ 12/19 \end{bmatrix}$$

$$4.4 (a) \quad P^n = \begin{bmatrix} \left(\frac{1}{2}\right)^n & 0 \\ 1 - \left(\frac{1}{2}\right)^n & 1 \end{bmatrix}, \quad n = 1, 2, \dots \text{ by induction.}$$

Thus no integer power of P has all positive entries.

$$(b) \quad P^n \rightarrow \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \text{ as } n \text{ increases, and so } P^n x^{(0)} \rightarrow \begin{bmatrix} 0 \\ 1 \end{bmatrix} \text{ for any } x^{(0)} \text{ as } n \text{ increases.}$$

(c) The entries of the limiting vector $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ are not all positive.

$$4.6 \quad P^2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} \text{ has all positive entries; } q = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$4.7 \quad 10/13$$

$$4.8 \quad 54\frac{1}{6}\% \text{ in region 1, } 16\frac{2}{3}\% \text{ in region 2, and } 29\frac{1}{6}\% \text{ in region 3.}$$

Chapter 5, page 63

$$5.1 (a) \quad \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad (b) \quad \begin{bmatrix} 6 \\ 5 \\ 6 \end{bmatrix}, \quad (c) \quad \begin{bmatrix} 78 \\ 54 \\ 79 \end{bmatrix}.$$

5.2 (a) Use Corollary 5.1; all row sums are less than one.

(b) Use Corollary 5.2; all column sums are less than one.

$$(c) \text{ Use Theorem 5.3 with } x = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} > Cx = \begin{bmatrix} 1.9 \\ .9 \\ .9 \end{bmatrix}.$$

5.3 E^2 has all positive entries.

5.4 Price of tomatoes: \$120.00. Price of corn: \$100.00. Price of lettuce: \$106.67.

5.5 \$1256 for the CE, \$1448 for the EE, \$1556 for the ME.

Chapter 6, page 78

6.1 The second class; \$15,000.

6.2 \$223

6.3 1 : 1.90 : 3.02 : 4.24 : 5.00

6.5 $s / (g_1^{-1} + g_2^{-1} + \dots + g_{k-1}^{-1})$

6.6 1 : 2 : 3 : \dots : n-1

Chapter 7, page 91

$$7.1 \text{ (a)} \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & 0 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 \end{bmatrix} \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix}$$

$$(b) \quad t = \begin{bmatrix} 1/4 \\ 3/4 \\ 1/4 \\ 3/4 \end{bmatrix}$$

$$(c) \quad t^{(1)} = \begin{bmatrix} 0 \\ \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix}, \quad t^{(2)} = \begin{bmatrix} 1/8 \\ 5/8 \\ 1/8 \\ 5/8 \end{bmatrix}, \quad t^{(3)} = \begin{bmatrix} 3/16 \\ 11/16 \\ 3/16 \\ 11/16 \end{bmatrix},$$

$$t^{(4)} = \begin{bmatrix} 7/32 \\ 23/32 \\ 7/32 \\ 23/32 \end{bmatrix}, \quad t^{(5)} = \begin{bmatrix} 15/64 \\ 47/64 \\ 15/64 \\ 47/64 \end{bmatrix}, \quad t^{(5)} - t = \begin{bmatrix} -1/64 \\ -1/64 \\ -1/64 \\ -1/64 \end{bmatrix}$$

(d) for t_1 : 5.4%; for t_2 : -1.7%.

7.2 $\frac{1}{2}$

$$7.3 \quad t^{(1)} = \left[\frac{3}{4} \quad \frac{5}{4} \quad \frac{2}{4} \quad \frac{5}{4} \quad \frac{4}{4} \quad \frac{2}{4} \quad \frac{5}{4} \quad \frac{4}{4} \quad \frac{3}{4} \right]^t$$

$$t^{(2)} = \left[\frac{13}{16} \quad \frac{18}{16} \quad \frac{9}{16} \quad \frac{22}{16} \quad \frac{13}{16} \quad \frac{7}{16} \quad \frac{21}{16} \quad \frac{16}{16} \quad \frac{10}{16} \right]^t$$

Chapter 8, page 106

$$8.2 \quad \begin{cases} a_n = \frac{1}{4} + \left(\frac{1}{2}\right)^{n+1} (a_0 - c_0) \\ b_n = \frac{1}{2} \\ c_n = \frac{1}{4} - \left(\frac{1}{2}\right)^{n+1} (a_0 - c_0) \end{cases} \quad n = 1, 2, \dots \quad \begin{cases} a_n \rightarrow \frac{1}{4} \\ b_n = \frac{1}{2} \\ c_n \rightarrow \frac{1}{4} \end{cases} \quad \text{as } n \rightarrow \infty$$

$$8.3 \quad \begin{cases} a_{2n+1} = \frac{2}{3} + \frac{1}{6(4)^n} (2a_0 - b_0 - 4c_0) \\ b_{2n+1} = \frac{1}{3} - \frac{1}{6(4)^n} (2a_0 - b_0 - 4c_0) \\ c_{2n+1} = 0 \end{cases} \quad n = 0, 1, 2, \dots$$

$$\begin{cases} a_{2n} = \frac{5}{12} + \frac{1}{6(4)^n} (2a_0 - b_0 - 4c_0) \\ b_{2n} = \frac{1}{2} \\ c_{2n} = \frac{1}{12} - \frac{1}{6(4)^n} (2a_0 - b_0 - 4c_0) \end{cases} \quad n = 1, 2, \dots$$

8.4 Eigenvalues: $\lambda_1 = 1$, $\lambda_2 = \frac{1}{2}$

Eigenvectors: $e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $e_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

8.5 12 generations; from 25% to .006%

$$8.6 \quad x^{(n)} = \begin{bmatrix} \frac{1}{2} + \frac{1}{2^{2n+3}} \left[(-3 - \sqrt{5})(1 + \sqrt{5})^{n+1} + (-3 + \sqrt{5})(1 - \sqrt{5})^{n+1} \right] \\ \frac{1}{2^{2n+1}} \left[(1 + \sqrt{5})^{n+1} + (1 - \sqrt{5})^{n+1} \right] \\ \frac{1}{2^{2n+1}} \left[(1 + \sqrt{5})^n + (1 - \sqrt{5})^n \right] \\ \frac{1}{2^{2n+1}} \left[(1 + \sqrt{5})^n + (1 - \sqrt{5})^n \right] \\ \frac{1}{2^{2n+1}} \left[(1 + \sqrt{5})^{n+1} + (1 - \sqrt{5})^{n+1} \right] \\ \frac{1}{2} + \frac{1}{2^{2n+3}} \left[(-3 - \sqrt{5})(1 + \sqrt{5})^{n+1} + (-3 + \sqrt{5})(1 - \sqrt{5})^{n+1} \right] \end{bmatrix}$$

$$x^{(n)} \rightarrow \begin{bmatrix} \frac{1}{2} \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{1}{2} \end{bmatrix} \text{ as } n \rightarrow \infty$$

$$8.8 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Chapter 9, page 122

$$9.1 \text{ (a) } \lambda_1 = 3/2, \quad x_1 = \begin{bmatrix} 1 \\ 1/3 \end{bmatrix}$$

$$\text{(b) } x^{(1)} = \begin{bmatrix} 100 \\ 50 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} 175 \\ 50 \end{bmatrix}, \quad x^{(3)} = \begin{bmatrix} 250 \\ 88 \end{bmatrix}, \quad x^{(4)} = \begin{bmatrix} 382 \\ 125 \end{bmatrix}$$

$$x^{(5)} = \begin{bmatrix} 570 \\ 191 \end{bmatrix}$$

$$\text{(c) } x^{(6)} = Lx^{(5)} = \begin{bmatrix} 857 \\ 285 \end{bmatrix}, \quad x^{(6)} \approx \lambda_1 x^{(5)} = \begin{bmatrix} 855 \\ 287 \end{bmatrix}$$

9.7 2.375

9.8 1.49611

Chapter 10, page 136

$$10.1 \text{ (a) Yield} = 33\frac{1}{3}\% \text{ of population; } x_1 = \begin{bmatrix} 1 \\ 1/3 \\ 1/18 \end{bmatrix}$$

$$\text{(b) Yield} = 45.8\% \text{ of population; } x_1 = \begin{bmatrix} 1 \\ 1/2 \\ 1/8 \end{bmatrix}; \text{ harvest } 57.9\% \text{ of youngest age class.}$$

$$10.2 \quad x_1 = \begin{bmatrix} 1.000 \\ .845 \\ .824 \\ .795 \\ .755 \\ .699 \\ .626 \\ .532 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad Lx_1 = \begin{bmatrix} 2.090 \\ .845 \\ .824 \\ .795 \\ .755 \\ .699 \\ .626 \\ .532 \\ .418 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \frac{1.090 + .418}{7.584} = .199$$

$$10.4 \quad h_I = (R-1) / (a_I b_1 b_2 \cdots b_{I-1} + \cdots + a_n b_1 b_2 \cdots b_{n-1})$$

$$10.5 \quad h_I = \frac{a_1 + a_2 b_1 + \cdots + a_{J-1} b_1 b_2 \cdots b_{J-2} - 1}{a_I b_1 b_2 \cdots b_{I-1} + \cdots + a_{J-1} b_1 b_2 \cdots b_{J-2}}$$

Chapter 11, page 146

$$11.1 \quad y = -\frac{1}{2} + \frac{7}{2}x$$

$$11.2 \quad y = \frac{2}{3} + \frac{1}{6}x$$

$$11.3 \quad y = 2 + 5x - 3x^2$$

$$11.4 \quad y = -5 + 3x - 4x^2 + 2x^3$$

$$11.9 \quad y = 4 - .2x + .2x^2; \text{ if } x = 12 \text{ then } y = 30.4 \text{ (\$30.4 thousand)}$$

Chapter 12, page 161

$$12.1 \quad \frac{\pi^2}{3} + 4\cos t + \cos 2t + \frac{4}{9}\cos 3t$$

$$12.2 \quad \frac{T^2}{3} + \frac{T^2}{\pi^2} \left(\cos \frac{2\pi}{T}t + \frac{1}{2^2} \cos \frac{4\pi}{T}t + \frac{1}{3^2} \cos \frac{6\pi}{T}t + \frac{1}{4^2} \cos \frac{8\pi}{T}t \right) \\ + \frac{T^2}{\pi} \left(\sin \frac{2\pi}{T}t + \frac{1}{2} \sin \frac{4\pi}{T}t + \frac{1}{3} \sin \frac{6\pi}{T}t + \frac{1}{4} \sin \frac{8\pi}{T}t \right)$$

$$12.3 \quad \frac{1}{\pi} + \frac{1}{2} \sin t - \frac{2}{3\pi} \cos 2t - \frac{2}{15\pi} \cos 4t$$

$$12.4 \frac{4}{\pi} \left(\frac{1}{2} - \frac{1}{1 \cdot 3} \cos t - \frac{1}{3 \cdot 5} \cos 2t - \frac{1}{5 \cdot 7} \cos 3t - \dots \right)$$

$$12.5 \frac{T}{4} - \frac{8T}{\pi^2} \left(\frac{1}{2^2} \cos \frac{2\pi t}{T} + \frac{1}{6^2} \cos \frac{6\pi t}{T} + \frac{1}{10^2} \cos \frac{10\pi t}{T} + \dots \right)$$

Chapter 13, page 177

13.1 $x_1 = 2$, $x_2 = 2/3$; maximum value of $z = 22/3$

13.2 no feasible solutions

13.3 unbounded solution

13.4 invest \$6,000 in bond A and \$4,000 in bond B; the annual yield is \$820.

13.5 25/36 cup of milk, 14/9 ounces of corn flakes; minimum cost = $655/36 = 18.2\bar{4}$

13.6 550 containers from company A and 300 containers from company B; maximum shipping charges = \$2,110

13.7 925 containers from company A and no containers from company B; maximum shipping charges = \$2,312.50

13.8 .4 pound of ingredient A and 2.4 pounds of ingredient B; minimum cost = 24.8¢

Chapter 14, page 197

14.1 Maximize $z' = -2x_1 - 5x_2$

subject to

$$3x_1 - 6x_2 + x_3 = 2$$

$$x_1 + x_2 + x_4 = 3$$

$$-x_1 + x_5 = -6$$

$$x_2 + x_6 = 5$$

and

$$x_1, x_2, x_3, x_4, x_5, x_6 \geq 0$$

14.2 Maximize $z = -3x_1 + x_2 + x_3$

subject to

$$3x_1 - 5x_2 + x_3 = 3$$

$$-2x_1 - x_2 + x_4 = 2$$

$$-x_1 + x_5 = -5$$

$$-x_2 + x_6 = -2$$

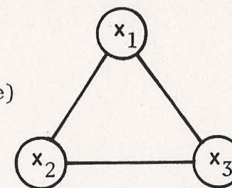
and

$$x_1, x_2, x_3, x_4, x_5, x_6 \geq 0$$

$$14.3 \text{ (b)} \quad \begin{bmatrix} 1/2 \\ 3/2 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 0 \\ 0 \\ -3 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 3/2 \\ 1/4 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 2 \\ 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ -3 \end{bmatrix}$$

$$\text{(c)} \quad x_1 = \begin{bmatrix} 1/2 \\ 3/2 \\ 0 \\ 0 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 0 \\ 3/2 \\ 1/4 \\ 0 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 0 \\ 2 \\ 0 \\ 1 \end{bmatrix}$$

$$\text{(d)} \quad x_1 = 0, \quad x_2 = 2, \quad x_3 = 0, \quad x_4 = 1; \quad z = 1$$



$$14.4 \text{ (a)} \quad \begin{bmatrix} -11/7 \\ 5/7 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} -1/7 \\ 0 \\ 5/7 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ -1/14 \\ 11/14 \end{bmatrix}$$

14.5 (a) Maximize $z = 3x_1 + 2x_2 - x_3 + 0x_4 + 0x_5$

subject to

$$2x_1 + 3x_2 + x_3 + x_4 = 4$$

$$x_1 + 2x_2 + 3x_3 + x_5 = 5$$

and

$$x_1, x_2, x_3, x_4, x_5 \geq 0$$

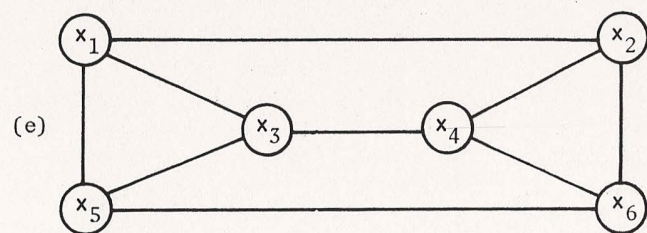
$$(b) \begin{bmatrix} -7 \\ 6 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 7/5 \\ 0 \\ 6/5 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 5 \\ 0 \\ 0 \\ -6 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \\ 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 5/2 \\ 0 \\ -7/2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 4/3 \\ 0 \\ 0 \\ 7/3 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 5/3 \\ 7/3 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 4 \\ 0 \\ -7 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 4 \\ 5 \end{bmatrix}$$

$$(c) x_1 = \begin{bmatrix} 7/5 \\ 0 \\ 6/5 \\ 0 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 2 \\ 0 \\ 0 \\ 0 \\ 3 \end{bmatrix}; x_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix},$$

$$x_4 = \begin{bmatrix} 0 \\ 4/3 \\ 0 \\ 0 \\ 7/3 \end{bmatrix}, x_5 = \begin{bmatrix} 0 \\ 0 \\ 5/3 \\ 7/3 \\ 0 \end{bmatrix}, x_6 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 4 \\ 5 \end{bmatrix}$$

$$(d) x_1 = 2, x_2 = 0, x_3 = 0, x_4 = 0, x_5 = 3; z = 6$$



$$(f) x_1 = 2, x_2 = 0, x_3 = 0; z = 6$$

14.6 (a) Maximize $z' = -2x_1 - 3x_2 - x_3 + 0x_4$

subject to

$$x_1 - 2x_2 + 3x_3 + x_4 = 5$$

$$2x_1 + x_2 - 2x_3 = 2$$

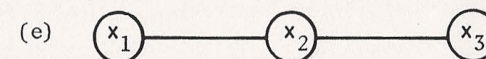
and

$$x_1, x_2, x_3, x_4 \geq 0$$

$$(b) \begin{bmatrix} 9/5 \\ -8/5 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 4 \end{bmatrix}, \begin{bmatrix} 0 \\ -16 \\ -9 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \\ 0 \\ 9 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ -1 \\ 8 \end{bmatrix}$$

$$(c) x_1 = \begin{bmatrix} 2 \\ 0 \\ 1 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 4 \end{bmatrix}, x_3 = \begin{bmatrix} 0 \\ 2 \\ 0 \\ 9 \end{bmatrix}$$

$$(d) x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 4; z' = -2$$



$$(f) x_1 = 1, x_2 = 0, x_3 = 0; z = 2$$

Chapter 15, page 218

15.1 $x_1 = 0, x_2 = 3/2; z = 6$

15.2 $x_1 = 2/3, x_2 = 1; z = 7/3$

15.3 $x_1 = 0, x_2 = 1/2, x_3 = 9/2; z = 26$

15.4 $x_1 = 22/3, x_2 = 6/13, x_3 = 0; z = 50/13$

15.5 $x_1 = 16/5, x_2 = 0, x_3 = 0, x_4 = 2/5; z = 10$

15.6 $x_1 = 0, x_2 = 13, x_3 = 34, x_4 = 0; z = 76$

15.7 $x_1 = 260, x_2 = 0; z = 520$

15.8 $x_1 = 2, x_2 = 2/3; z = 22/3$

15.9 $x_1 = 550, x_2 = 300; z = 2110$

15.10 $x_1 = 925, x_2 = 0; z = 2312.50$