NONLINEAR PROGRAMMING

Theory and Algorithms

Third Edition

MOKHTAR S. BAZARAA
Georgia Institute of Technology
School of Industrial and Systems Engineering
Atlanta, Georgia

HANIF D. SHERALI
Virginia Polytechnic Institute and State University
Grado Department of Industrial and Systems Engineering
Blacksburg, Virginia

C. M. SHETTY
Georgia Institute of Technology
School of Industrial and Systems Engineering
Atlanta, Georgia

A JOHN WILEY & SONS, INC., PUBLICATION

This Page Intentionally Left Blank

NONLINEAR PROGRAMMING

This Page Intentionally Left Blank

NONLINEAR PROGRAMMING

Theory and Algorithms

Third Edition

MOKHTAR S. BAZARAA
Georgia Institute of Technology
School of Industrial and Systems Engineering
Atlanta, Georgia

HANIF D. SHERALI
Virginia Polytechnic Institute and State University
Grado Department of Industrial and Systems Engineering
Blacksburg, Virginia

C. M. SHETTY
Georgia Institute of Technology
School of Industrial and Systems Engineering
Atlanta, Georgia

A JOHN WILEY & SONS, INC., PUBLICATION

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

Library of Congress Cataloging-in-Publication Data:

Bazaraa, M. S.
Nonlinear programming : theory and algorithms / Mokhtar S. Bazaraa, Hanif D. Sherali, C. M. Shetty.—3rd ed.
p. cm.
"Wiley-Interscience."
Includes bibliographical references and index.
ISBN-13: 978-0-471-48600-8 (cloth: alk. paper)
ISBN-10: 0-471-48600-0 (cloth: alk. paper)
1. Nonlinear programming. I. Sherali, Hanif D., 1952–. II. Shetty, C. M., 1929–. III Title.

T57.8.B39 2006
519.7'6—dc22 2005054230

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Dedicated to our parents

This Page Intentionally Left Blank

Contents

This Page Intentionally Left Blank

Preface

Nonlinear programming deals with the problem of optimizing an objective function in the presence of equality and inequality constraints. If all the functions are linear, we obviously have a *linear program.* Otherwise, the problem is called a *nonlinear program.* The development of highly efficient and robust algorithms and software for linear programming, the advent of high-speed computers, and the education of managers and practitioners in regard to the advantages and profitability of mathematical modeling and analysis have made linear programming an important tool for solving problems in diverse fields. However, many realistic problems cannot be adequately represented or approximated as a linear program, owing to the nature of the nonlinearity of the objective function and/or the nonlinearity of any of the constraints. Efforts to solve such nonlinear problems efficiently have made rapid progress during the past four decades. This book presents these developments in a logical and self-contained form.

The book is divided into three major parts dealing, respectively, with convex analysis, optimality conditions and duality, and computational methods. Convex analysis involves convex sets and convex functions and is central to the study of the field of optimization. The ultimate goal in optimization studies is to develop efficient computational schemes for solving the problem at hand. Optimality conditions and duality can be used not only to develop termination criteria but also to motivate and design the computational method itself.

In preparing this book, a special effort has been made to make certain that it is self-contained and that it is suitable both as a text and as a reference. Within each chapter, detailed numerical examples and graphical illustrations have been provided to aid the reader in understanding the concepts and methods discussed. In addition, each chapter contains many exercises. These include (1) simple numerical problems to reinforce the material discussed in the text, (2) problems introducing new material related to that developed in the text, and (3) theoretical exercises meant for advanced students. At the end of each chapter, extensions, references, and material related to that covered in the text are presented. These notes should be useful to the reader for further study. The book also contains an extensive bibliography.

Chapter 1 gives several examples of problems from different engineering disciplines that can be viewed as nonlinear programs. Problems involving optimal control, both discrete and continuous, are discussed and illustrated by examples from production, inventory control, and highway design. Examples of a two-bar truss design and a two-bearing journal design are given. Steady-state conditions of an electrical network are discussed from the point of view of

obtaining an optimal solution to a quadratic program. A large–scale nonlinear model arising in the management of water resources is developed, and nonlinear models arising in stochastic programming and in location theory are discussed. Finally, we provide an important discussion on modeling and on formulating nonlinear programs from the viewpoint of favorably influencing the performance of algorithms that will ultimately be used for solving them.

The remaining chapters are divided into three parts. Part 1, consisting of Chapters 2 and 3, deals with convex sets and convex functions. Topological properties of convex sets, separation and support of convex sets, polyhedral sets, extreme points and extreme directions of polyhedral sets, and linear programming are discussed in Chapter 2. Properties of convex functions, including subdifferentiability and minima and maxima over a convex set, are discussed in Chapter 3. Generalizations of convex functions and their interrelationships are also included, since nonlinear programming algorithms suitable for convex functions can be used for a more general class involving pseudoconvex and quasiconvex functions. The appendix provides additional tests for checking generalized convexity properties, and we discuss the concept of convex envelopes and their uses in global optimization methods through the exercises.

Part 2, which includes Chapters 4 through 6, covers optimality conditions and duality. In Chapter 4, the classical Fritz John (FJ) and the Karush–Kuhn–Tucker (KKT) optimality conditions are developed for both inequality- and equality-constrained problems. First- and second-order optimality conditions are derived and higher-order conditions are discussed along with some cautionary examples. The nature, interpretation, and value of FJ and KKT points are also described and emphasized. Some foundational material on both first- and second-order constraint qualifications is presented in Chapter 5. We discuss interrelationships between various proposed constraint qualifications and provide insights through many illustrations. Chapter 6 deals with Lagrangian duality and saddle point optimality conditions. Duality theorems, properties of the dual function, and both differentiable and nondifferentiable methods for solving the dual problem are discussed. We also derive necessary and sufficient conditions for the absence of a duality gap and interpret this in terms of a suitable perturbation function. In addition, we relate Lagrangian duality to other special forms of duals for linear and quadratic programming problems. Besides Lagrangian duality, there are several other duality formulations in nonlinear programming, such as conjugate duality, min–max duality, surrogate duality, composite Lagrangian and surrogate duality, and symmetric duality. Among these, the Lagrangian duality seems to be the most promising in the areas of theoretical and algorithmic developments. Moreover, the results that can be obtained via these alternative duality formulations are closely related. In view of this, and for brevity, we have elected to discuss Lagrangian duality in the text and to introduce other duality formulations only in the exercises.

Part 3, consisting of Chapters 7 through 11, presents algorithms for solving both unconstrained and constrained nonlinear programming problems. Chapter 7 deals exclusively with convergence theorems, viewing algorithms as point-to-set maps. These theorems are used actively throughout the remainder of

the book to establish the convergence of the various algorithms. Likewise, we discuss the issue of rates of convergence and provide a brief discussion on criteria that can be used to evaluate algorithms.

Chapter 8 deals with the topic of unconstrained optimization. To begin, we discuss several methods for performing both exact and inexact line searches, as well as methods for minimizing a function of several variables. Methods using both derivative and derivative-free information are presented. Newton's method and its variants based on trust region and the Levenberg–Marquardt approaches are discussed. Methods that are based on the concept of conjugacy are also covered. In particular, we present quasi-Newton (variable metric) and conjugate gradient (fixed metric) algorithms that have gained a great deal of popularity in practice. We also introduce the subject of subgradient optimization methods for nondifferentiable problems and discuss variants fashioned in the spirit of conjugate gradient and variable metric methods. Throughout, we address the issue of convergence and rates of convergence for the various algorithms, as well as practical implementation aspects.

In Chapter 9 we discuss penalty and barrier function methods for solving nonlinear programs, in which the problem is essentially solved as a sequence of unconstrained problems. We describe general exterior penalty function methods, as well as the particular exact absolute value and the augmented Lagrangian penalty function approaches, along with the method of multipliers. We also present interior barrier function penalty approaches. In all cases, implementation issues and convergence rate characteristics are addressed. We conclude this chapter by describing a polynomial-time primal-dual path-following algorithm for linear programming based on a logarithmic barrier function approach. This method can also be extended to solve convex quadratic programs polynomially. More computationally effective *predictor–corrector* variants of this method are also discussed.

Chapter 10 deals with the method of feasible directions, in which, given a feasible point, a feasible improving direction is first found and then a new, improved feasible point is determined by minimizing the objective function along that direction. The original methods proposed by Zoutendijk and subsequently modified by Topkis and Veinott to assure convergence are presented. This is followed by the popular successive linear and quadratic programming approaches, including the use of ℓ_1 penalty functions either directly in the direction-finding subproblems or as merit functions to assure global convergence. Convergence rates and the Maratos effect are also discussed. This chapter also describes the gradient projection method of Rosen along with its convergent variants, the reduced gradient method of Wolfe and the generalized reduced gradient method, along with its specialization to Zangwill's convex simplex method. In addition, we unify and extend the reduced gradient and the convex simplex methods through the concept of suboptimization and the superbasic–basic–nonbasic partitioning scheme. Effective first- and second-order variants of this approach are discussed.

Finally, Chapter 11 deals with some special problems that arise in different applications as well as in the solution of other nonlinear programming problems. In particular, we present the linear complementary, quadratic

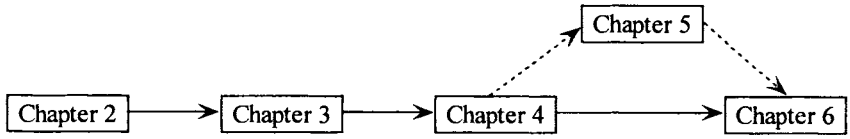
separable, linear fractional, and geometric programming problems. Methodologies used for solving these problems, such as the use of Lagrangian duality concepts in the algorithmic development for geometric programs, serve to strengthen the ideas described in the preceding chapters. Moreover, in the context of solving nonconvex quadratic problems, we introduce the concept of the *reformulation–linearization/convexification technique* (RLT) as a *global optimization* methodology for finding an optimal solution. The RLT can also be applied to general nonconvex polynomial and factorable programming problems to determine global optimal solutions. Some of these extensions are pursued in the exercises in Chapter 11. The Notes and References section provides directions for further study.

This book can be used both as a reference for topics in nonlinear programming and as a text in the fields of operations research, management science, industrial engineering, applied mathematics, and in engineering disciplines that deal with analytical optimization techniques. The material discussed requires some mathematical maturity and a working knowledge of linear algebra and calculus. For the convenience of the reader, Appendix A summarizes some mathematical topics used frequently in the book, including matrix factorization techniques.

As a text, the book can be used (1) in a course on foundations of optimization and (2) in a course on computational methods as detailed below. It can also be used in a two-course sequence covering all the topics.

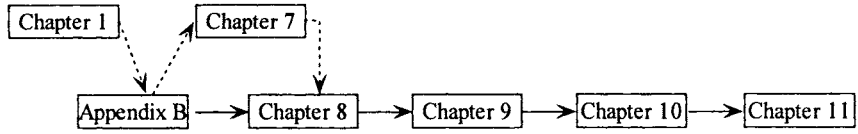
1. *Foundations of Optimization*

This course is meant for undergraduate students in applied mathematics and for graduate students in other disciplines. The suggested coverage is given schematically below, and it can be covered in the equivalent of a one-semester course. Chapter 5 could be omitted without loss of continuity. A reader familiar with linear programming may also skip Section 2.7.



2. *Computational Methods in Nonlinear Programming*

This course is meant for graduate students who are interested in algorithms for solving nonlinear programs. The suggested coverage is given schematically below, and it can be covered in the equivalent of a one-semester course. The reader who is not interested in convergence analyses may skip Chapter 7 and the discussion related to convergence in Chapters 8 through 11. The minimal background on convex analysis and optimality conditions needed to study Chapters 8 through 11 is summarized in Appendix B for the convenience of the reader. Chapter 1, which gives many examples of nonlinear programming problems, provides a good introduction to the course, but no continuity will be lost if this chapter is skipped.



Acknowledgements

We again express our thanks to Dr. Robert N. Lehrer, former director of the School of Industrial and Systems Engineering at the Georgia Institute of Technology, for his support in the preparation of the first edition of this book; to Dr. Jamie J. Goode of the School of Mathematics, Georgia Institute of Technology, for his friendship and active cooperation; and to Mrs. Carolyn Piersma, Mrs. Joene Owen, and Ms. Kaye Watkins for their typing of the first edition of this book.

In the preparation of the second edition of this book, we thank Professor Robert D. Dryden, head of the Department of Industrial and Systems Engineering at Virginia Polytechnic Institute and State University, for his support. We thank Dr. Gyunghyun Choi, Dr. Ravi Krishnamurthy, and Mrs. Semeen Sherali for their typing efforts, and Dr. Joanna Leleno for her diligent preparation of the (partial) solutions manual.

We thank Professor G. Don Taylor, head of the Department of Industrial and Systems Engineering at Virginia Polytechnic Institute and State University, for his support during the preparation of the present edition of the book. We also acknowledge the *National Science Foundation*, Grant Number 0094462, for supporting research on nonconvex optimization that is covered in Chapter 11. This edition was typed from scratch, including figures and tables, by Ms. Sandy Dalton. We thank her immensely for her painstaking and diligent effort at accomplishing this formidable task. We also thank Dr. Barbara Fraticelli for her insightful comments and laboriously careful reading of the manuscript.

Mokhtar S. Bazaraa
Hanif D. Sherali
C. M. Shetty

This Page Intentionally Left Blank

Chapter 1 Introduction

Operations research analysts, engineers, managers, and planners are traditionally confronted by problems that need solving. The problems may involve arriving at an optimal design, allocating scarce resources, planning industrial operations, or finding the trajectory of a rocket. In the past, a wide range of solutions was considered acceptable. In engineering design, for example, it was common to include a large safety factor. However, because of continued competition, it is no longer adequate to develop only an acceptable design. In other instances, such as in space vehicle design, the acceptable designs themselves may be limited. Hence, there is a real need to answer such questions as: Are we making the most effective use of our scarce resources? Can we obtain a more economical design? Are we taking risks within acceptable limits? In response to an ever-enlarging domain of such inquiries, there has been a very rapid growth of optimization models and techniques. Fortunately, the parallel growth of faster and more accurate sophisticated computing facilities has aided substantially in the use of the techniques developed.

Another aspect that has stimulated the use of a systematic approach to problem solving is the rapid increase in the size and complexity of problems as a result of the technological growth since World War II. Engineers and managers are called upon to study all facets of a problem and their complicated interrelationships. Some of these interrelationships may not even be well understood. Before a system can be viewed as a whole, it is necessary to understand how the components of the system interact. Advances in the techniques of measurement, coupled with statistical methods to test hypotheses, have aided significantly in this process of studying the interaction between components of the system.

The acceptance of the field of operations research in the study of industrial, business, military, and governmental activities can be attributed, at least in part, to the extent to which the operations research approach and methodology have aided the decision makers. Early postwar applications of operations research in the industrial context were mainly in the area of linear programming and the use of statistical analyses. Since that time, efficient procedures and computer codes have been developed to handle such problems. This book is concerned with nonlinear programming, including the characterization of optimal solutions and the development of algorithmic procedures.

In this chapter we introduce the nonlinear programming problem and discuss some simple situations that give rise to such a problem. Our purpose is only to provide some background on nonlinear problems; indeed, an exhaustive

discussion of potential applications of nonlinear programming can be the subject matter of an entire book. We also provide some guidelines here for constructing models and problem formulations from the viewpoint of enhancing algorithmic efficiency and problem solvability. Although many of these remarks will be better appreciated as the reader progresses through the book, it is best to bear these general fundamental comments in mind at the very onset.

1.1 Problem Statement and Basic Definitions

Consider the following nonlinear programming problem:

$$\begin{aligned} \text{Minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0 \quad \text{for } i = 1,...,m \\ & h_i(\mathbf{x}) = 0 \quad \text{for } i = 1,...,\ell \\ & \mathbf{x} \in X, \end{aligned}$$

where f, $g_1,...,g_m$, $h_1,...,h_\ell$ are functions defined on R^n, X is a subset of R^n, and $\mathbf{x}$ is a vector of n components $x_1,..., x_n$. The above problem must be solved for the values of the variables $x_1,..., x_n$ that satisfy the restrictions and meanwhile minimize the function f.

The function f is usually called the *objective function,* or the *criterion function.* Each of the constraints $g_i(\mathbf{x}) \leq 0$ for $i = 1,..., m$ is called an *inequality constraint*, and each of the constraints $h_i(\mathbf{x}) = 0$ for $i = 1,...,\ell$ is called an *equality constraint.* The set X might typically include lower and upper bounds on the variables, which even if implied by the other constraints can play a useful role in some algorithms. Alternatively, this set might represent some specially structured constraints that are highlighted to be exploited by the optimization routine, or it might represent certain regional containment or other complicating constraints that are to be handled separately via a special mechanism. A vector $\mathbf{x} \in X$ satisfying all the constraints is called a *feasible solution* to the problem. The collection of all such solutions forms the *feasible region.* The nonlinear programming problem, then, is to find a feasible point $\bar{\mathbf{x}}$ such that $f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$ for each feasible point $\mathbf{x}$. Such a point $\bar{\mathbf{x}}$ is called an *optimal solution*, or simply a *solution*, to the problem. If more than one optimum exists, they are referred to collectively as *alternative optimal solutions.*

Needless to say, a nonlinear programming problem can be stated as a maximization problem, and the inequality constraints can be written in the form $g_i(\mathbf{x}) \geq 0$ for $i = 1,..., m$. In the special case when the objective function is linear and when all the constraints, including the set X, can be represented by linear inequalities and/or linear equations, the above problem is called a *linear program.*

To illustrate, consider the following problem:

$$\begin{aligned} \text{Minimize} \quad & (x_1-3)^2+(x_2-2)^2 \\ \text{subject to} \quad & x_1^2-x_2-3\le 0 \\ & x_2-1\le 0 \\ & -x_1\le 0. \end{aligned}$$

The objective function and the three inequality constraints are

$$\begin{aligned} f(x_1,x_2) &= (x_1-3)^2+(x_2-2)^2 \\ g_1(x_1,x_2) &= x_1^2-x_2-3 \\ g_2(x_1,x_2) &= x_2-1 \\ g_3(x_1,x_2) &= -x_1. \end{aligned}$$

Figure 1.1 illustrates the feasible region. The problem, then, is to find a point in the feasible region having the smallest possible value of $(x_1-3)^2+(x_2-2)^2$. Note that points (x_1,x_2) with $(x_1-3)^2+(x_2-2)^2=c$ represent a circle with radius $\sqrt{c}$ and center (3, 2). This circle is called the *contour* of the objective function having the value c. Since we wish to minimize f, we must find the contour circle having the smallest radius that intersects the feasible region. As shown in Figure 1.1, the smallest such circle has $c = 2$ and intersects the feasible region at the point (2, 1). Therefore, the optimal solution occurs at the point (2, 1) and has an objective value equal to 2.

The approach used above is to find an optimal solution by determining the objective contour having the smallest objective value that intersects the feasible region. Obviously, this approach of solving the problem geometrically is only suitable for small problems and is not practical for problems having more than two variables or those having complicated objective and constraint functions.

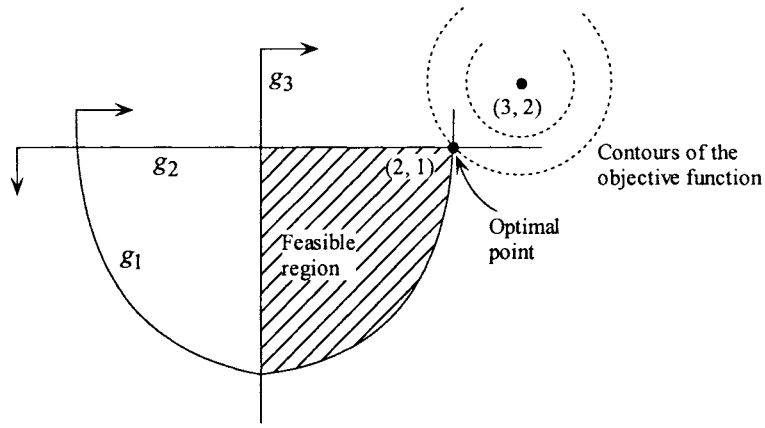


Figure 1.1 Geometric solution of a nonlinear problem.

Notation

The following notation is used throughout the book. Vectors are denoted by boldface lowercase Roman letters, such as **x**, **y**, and **z**. All vectors are column vectors unless stated explicitly otherwise. Row vectors are the transpose of column vectors; for example, $\mathbf{x}^t$ denotes the row vector $(x_1,..., x_n)$. The n-dimensional *real Euclidean space*, composed of all real vectors of dimension n, is denoted by R^n. Matrices are denoted by boldface capital Roman letters, such as **A** and **B**. Scalar-valued functions are denoted by lowercase Roman or Greek letters, such as f, g, and θ. Vector-valued functions are denoted by boldface lowercase Roman or Greek letters, such as **g** and Ψ. Point-to-set maps are denoted by boldface capital Roman letters such as **A** and **B**. Scalars are denoted by lowercase Roman and Greek letters, such as k, λ, and α.

1.2 Illustrative Examples

In this section we discuss some example problems that can be formulated as nonlinear programs. In particular, we discuss optimization problems in the following areas:

A. Optimal control
B. Structural design
C. Mechanical design
D. Electrical networks
E. Water resources management
F. Stochastic resource allocation
G. Location of facilities

A. Optimal Control Problems

As we shall learn shortly, a discrete control problem can be stated as a nonlinear programming problem. Furthermore, a continuous optimal control problem can be approximated by a nonlinear programming problem. Hence, the procedures discussed later in the book can be used to solve some optimal control problems.

Discrete Optimal Control

Consider a fixed-time discrete optimal control problem of duration K periods. At the beginning of period k, the system is represented by the *state vector* $\mathbf{y}_{k-1}$. A *control vector* $\mathbf{u}_k$ changes the state of the system from $\mathbf{y}_{k-1}$ to $\mathbf{y}_k$ at the end of period k according to the following relationship:

$$\mathbf{y}_k = \mathbf{y}_{k-1} + \boldsymbol{\phi}_k(\mathbf{y}_{k-1}, \mathbf{u}_k) \qquad \text{for } k = 1,..., K.$$

Given the initial state $\mathbf{y}_0$, applying the sequence of controls $\mathbf{u}_1,...,\mathbf{u}_K$ would result in a sequence of state vectors $\mathbf{y}_1,...,\mathbf{y}_K$ called the *trajectory*. This process is illustrated in Figure 1.2.

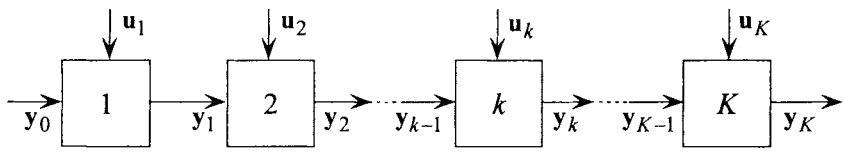


Figure 1.2 Discrete control system.

A sequence of controls $\mathbf{u}_1,\ldots,\mathbf{u}_K$ and a sequence of state vectors $\mathbf{y}_0$, $\mathbf{y}_1,\ldots,\mathbf{y}_K$ are called *admissible* or *feasible* if they satisfy the following restrictions:

$$\begin{aligned} &\mathbf{y}_k \in Y_k && \text{for } k = 1,\ldots,K \\ &\mathbf{u}_k \in U_k && \text{for } k = 1,\ldots,K \\ &\Psi(\mathbf{y}_0,\ldots,\mathbf{y}_K,\mathbf{u}_1,\ldots,\mathbf{u}_K) \in D, \end{aligned}$$

where $Y_1,\ldots,Y_K$, $U_1,\ldots,U_K$, and D are specified sets, and Ψ is a known function, usually called the *trajectory constraint function.* Among all feasible controls and trajectories, we seek a control and a corresponding trajectory that optimize a certain objective function. The discrete control problem can thus be stated as follows:

$$\begin{aligned} \text{Minimize} \quad & \alpha(\mathbf{y}_0,\mathbf{y}_1,\ldots,\mathbf{y}_K,\mathbf{u}_1,\ldots,\mathbf{u}_K) \\ \text{subject to} \quad & \mathbf{y}_k = \mathbf{y}_{k-1} + \phi_k(\mathbf{y}_{k-1},\mathbf{u}_k) && \text{for } k = 1,\ldots,K \\ & \mathbf{y}_k \in Y_k && \text{for } k = 1,\ldots,K \\ & \mathbf{u}_k \in U_k && \text{for } k = 1,\ldots,K \\ & \Psi(\mathbf{y}_0,\ldots,\mathbf{y}_K,\mathbf{u}_1,\ldots,\mathbf{u}_K) \in D. \end{aligned}$$

Combining $\mathbf{y}_1,\ldots,\mathbf{y}_K$, $\mathbf{u}_1,\ldots,\mathbf{u}_K$ as the vector $\mathbf{x}$, and by suitable choices of $\mathbf{g}$, $\mathbf{h}$, and X, it can easily be verified that the above problem can be stated as the nonlinear programming problem introduced in Section 1.1.

Production-Inventory Example We illustrate the formulation of a *discrete control problem* with the following production-inventory example. Suppose that a company produces a certain item to meet a known demand, and suppose that the production schedule must be determined over a total of K periods. The demand during any period can be met from the inventory at the beginning of the period and the production during the period. The maximum production during any period is restricted by the production capacity of the available equipment so that it cannot exceed b units. Assume that adequate temporary labor can be hired when needed and laid off if superfluous. However, to discourage heavy labor fluctuations, a cost proportional to the square of the difference in the labor force during any two successive periods is assumed. Also, a cost proportional to the inventory carried forward from one period to another is

incurred. Find the labor force and inventory during periods 1,..., K such that the demand is satisfied and the total cost is minimized.

In this problem, there are two state variables, the inventory level I_k and the labor force L_k at the end of period k. The control variable u_k is the labor force acquired during period k ($u_k < 0$ means that the labor is reduced by an amount $-u_k$). The production-inventory problem can thus be stated as follows:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{k=1}^{K} (c_1 u_k^2 + c_2 I_k) \\
\text{subject to} \quad & L_k = L_{k-1} + u_k && \text{for } k = 1,\ldots,K \\
& I_k = I_{k-1} + pL_{k-1} - d_k && \text{for } k = 1,\ldots,K \\
& 0 \le L_k \le b/p && \text{for } k = 1,\ldots,K \\
& I_k \ge 0 && \text{for } k = 1,\ldots,K,
\end{aligned}$$

where the initial inventory I_0 and the initial labor force L_0 are known, d_k is the known demand during period k, and p is the number of units produced per worker during any given period.

Continuous Optimal Control

In the case of a discrete control problem, the controls are exercised at discrete points. We now consider a *fixed-time continuous control problem* in which a control function, $\mathbf{u}$, is to be exerted over the planning horizon [0, T]. Given the initial state $\mathbf{y}_0$, the relationship between the state vector $\mathbf{y}$ and the control vector $\mathbf{u}$ is governed by the following differential equation:

$$\dot{\mathbf{y}}(t) = \phi[\mathbf{y}(t), \mathbf{u}(t)] \qquad \text{for } t \in [0,T].$$

The control function and the corresponding trajectory function are called *admissible* if the following restrictions hold true:

$$\begin{aligned}
& \mathbf{y}(t) \in Y && \text{for } t \in [0,T] \\
& \mathbf{u}(t) \in U && \text{for } t \in [0,T] \\
& \Psi(\mathbf{y},\mathbf{u}) \in D.
\end{aligned}$$

A typical example of the set U is the collection of piecewise continuous functions on [0, T] such that $\mathbf{a} \le \mathbf{u}(t) \le \mathbf{b}$ for $t \in [0,T]$. The optimal control problem can be stated as follows, where the initial state vector $\mathbf{y}(0) = \mathbf{y}_0$ is given:

$$\begin{aligned}
\text{Minimize} \quad & \int_0^T \alpha[\mathbf{y}(t), \mathbf{u}(t)]\, dt \\
\text{subject to} \quad & \dot{\mathbf{y}}(t) = \phi[\mathbf{y}(t), \mathbf{u}(t)] && \text{for } t \in [0,T] \\
& \mathbf{y}(t) \in Y && \text{for } t \in [0,T] \\
& \mathbf{u}(t) \in U && \text{for } t \in [0,T] \\
& \Psi(\mathbf{y},\mathbf{u}) \in D.
\end{aligned}$$

A continuous optimal control problem can be approximated by a discrete problem. In particular, suppose that the planning region $[0, T]$ is divided into K periods, each of duration Δ, such that $K\Delta = T$. Denoting $\mathbf{y}(k\Delta)$ by $\mathbf{y}_k$ and $\mathbf{u}(k\Delta)$ by $\mathbf{u}_k$, for $k = 1, ..., K$, the above problem can be approximated as follows, where the initial state $\mathbf{y}_0$ is given:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{k=1}^{K} \alpha(\mathbf{y}_k, \mathbf{u}_k) \\
\text{subject to} \quad & \mathbf{y}_k = \mathbf{y}_{k-1} + \Delta\boldsymbol{\phi}(\mathbf{y}_{k-1}, \mathbf{u}_k) && \text{for } k = 1,..., K \\
& \mathbf{y}_k \in Y && \text{for } k = 1,..., K \\
& \mathbf{u}_k \in U && \text{for } k = 1,..., K \\
& \Psi(\mathbf{y}_0,..., \mathbf{y}_K, \mathbf{u}_1, \ldots, \mathbf{u}_K) \in D.
\end{aligned}$$

Example of Rocket Launching Consider the problem of a rocket that is to be moved from ground level to a height $\bar{y}$ in time T. Let $y(t)$ denote the height from the ground at time t, and let $u(t)$ denote the force exerted in the vertical direction at time t. Assuming that the rocket has mass m, the equation of motion is given by

$$m\ddot{y}(t) + mg = u(t) \qquad \text{for } t \in [0, T],$$

where $\ddot{y}(t)$ is the acceleration at time t and g is the deceleration due to gravity. Furthermore, suppose that the maximum force that could be exerted at any time cannot exceed b. If the objective is to expend the smallest possible energy so that the rocket reaches an altitude $\bar{y}$ at time T, the problem can be formulated as follows:

$$\begin{aligned}
\text{Minimize} \quad & \int_0^T |u(t)|\, \dot{y}(t)\, dt \\
\text{subject to} \quad & m\ddot{y}(t) + mg = u(t) && \text{for } t \in [0, T] \\
& |u(t)| \le b && \text{for } t \in [0, T] \\
& y(T) = \bar{y},
\end{aligned}$$

where $y(0) = 0$. This problem having a second-order differential equation can be transformed into an equivalent problem having two first-order differential equations. This can be done by the following substitution: $y_1 = y$ and $y_2 = \dot{y}$. Therefore, $m\ddot{y} + mg = u$ is equivalent to $\dot{y}_1 = y_2$ and $m\dot{y}_2 + mg = u$. Hence, the problem can be restated as follows:

$$\begin{aligned}
\text{Minimize} \quad & \int_0^T |u(t)|\, y_2(t)\, dt \\
\text{subject to} \quad & \dot{y}_1(t) = y_2(t) && \text{for } t \in [0, T] \\
& m\dot{y}_2(t) = u(t) - mg && \text{for } t \in [0, T] \\
& |u(t)| \le b && \text{for } t \in [0, T] \\
& y_1(T) = \bar{y},
\end{aligned}$$

where $y_1(0) = y_2(0) = 0$. Suppose that we divide the interval [0, T] into K periods. To simplify the notation, suppose that each period has length ℓ. Denoting the force, altitude, and velocity at the end of period k by u_k, $y_{1,k}$, and $y_{2,k}$, respectively, for $k = 1,..., K$, the above problem can be approximated by the following nonlinear program, where $y_{1,0} = y_{2,0} = 0$:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{k=1}^{K} |u_k| y_{2,k} \\
\text{subject to} \quad & y_{1,k} - y_{1,k-1} = y_{2,k-1} && \text{for } k = 1,..., K \\
& m(y_{2,k} - y_{2,k-1}) = u_k - mg && \text{for } k = 1,..., K \\
& |u_k| \le b && \text{for } k = 1,..., K \\
& y_{1,K} = \bar{y}.
\end{aligned}$$

The interested reader may refer to Luenberger [1969, 1973a/1984] for this problem and other continuous optimal control problems.

Example of Highway Construction Suppose that a road is to be constructed over uneven terrain. The construction cost is assumed to be proportional to the amount of dirt added or removed. Let T be the length of the road, and let $c(t)$ be the known height of the terrain at any given $t \in [0,T]$. The problem is to formulate an equation describing the height of the road $y(t)$ for $t \in [0,T]$.

To avoid excessive slopes on the road, the maximum slope must not exceed b_1 in magnitude; that is, $|\dot{y}(t)| \le b_1$. In addition, to reduce the roughness of the ride, the rate of change of the slope of the road must not exceed b_2 in magnitude; that is, $|\ddot{y}(t)| \le b_2$. Furthermore, the end conditions $y(0) = a$ and $y(T) = b$ must be observed. The problem can thus be stated as follows:

$$\begin{aligned}
\text{Minimize} \quad & \int_0^T |y(t) - c(t)|\, dt \\
\text{subject to} \quad & |\dot{y}(t)| \le b_1 && \text{for } t \in [0,T] \\
& |\ddot{y}(t)| \le b_2 && \text{for } t \in [0,T] \\
& y(0) = a \\
& y(T) = b.
\end{aligned}$$

Note that the control variable is the amount of dirt added or removed; that is, $u(t) = y(t) - c(t)$.

Now let $y_1 = y$ and $y_2 = \dot{y}$, and divide the road length into K intervals. For simplicity, suppose that each interval has length ℓ. Denoting $c(k)$, $y_1(k)$, and $y_2(k)$, by c_k, $y_{1,k}$, and $y_{2,k}$, respectively, the above problem can be approximated by the following nonlinear program:

$$\text{Minimize} \quad \sum_{k=1}^{K} |y_{1,k} - c_k|$$

$$\begin{aligned}
\text{subject to} \quad & y_{1,k} - y_{1,k-1} = y_{2,k-1} && \text{for } k = 1,\ldots,K \\
& -b_1 \le y_{2,k} \le b_1 && \text{for } k = 0,\ldots,K-1 \\
& -b_2 \le y_{2,k} - y_{2,k-1} \le b_2 && \text{for } k = 1,\ldots,K-1 \\
& y_{1,0} = a \\
& y_{1,K} = b.
\end{aligned}$$

The interested reader may refer to Citron [1969] for more details of this example.

B. Structural Design

Structural designers have traditionally endeavored to develop designs that could safely carry the projected loads. The concept of optimality was implicit only through the standard practice and experience of the designer. Recently, the design of sophisticated structures, such as aerospace structures, has called for more explicit consideration of optimality.

The main approaches used for minimum weight design of structural systems are based on the use of mathematical programming or other rigorous numerical techniques combined with structural analysis methods. Linear programming, nonlinear programming, and Monte Carlo simulation have been the principal techniques used for this purpose.

As noted by Batt and Gellatly [1974]:

> The total process for the design of a sophisticated aerospace structure is a multistage procedure that ranges from consideration of overall systems performance down to the detailed design of individual components. While all levels of the design process have some greater or lesser degree of interaction with each other, the past state-of-the-art in design has demanded the assumption of a relatively loose coupling between the stages. Initial work in structural optimization has tended to maintain this stratification of design philosophy, although this state of affairs has occurred, possibly, more as a consequence of the methodology used for optimization than from any desire to perpetuate the delineations between design stages.

The following example illustrates how structural analysis methods can be used to yield a nonlinear programming problem involving a minimum-weight design of a two-bar truss.

Two-Bar Truss Consider the planar truss shown in Figure 1.3. The truss consists of two steel tubes pinned together at one end and fixed at two pivot points at the other end. The span, that is, the distance between the two pivots, is fixed at $2s$. The design problem is to choose the height of the truss and

the thickness and average diameter of the steel tubes so that the truss will support a load of $2W$ while minimizing the total weight of the truss.

Denote the average tube diameter, tube thickness, and truss height by x_1, x_2, and x_3, respectively. The weight of the steel truss is then given by $2\pi\rho x_1 x_2 (s^2 + x_3^2)^{1/2}$, where ρ is the density of the steel tube. The following constraints must be observed:

1. Because of space limitations, the height of the truss must not exceed b_1; that is, $x_3 \le b_1$.
2. The ratio of the diameter of the tube to the thickness of the tube must not exceed b_2; that is, $x_1/x_2 \le b_2$.
3. The compression stress in the steel tubes must not exceed the steel yield stress. This gives the following constraint, where b_3 is a constant:

$$W(s^2 + x_3^2)^{1/2} \le b_3 x_1 x_2 x_3.$$

4. The height, diameter, and thickness must be chosen such that the tubes will not buckle under the load. This constraint can be expressed mathematically as follows, where b_4 is a known parameter:

$$W(s^2 + x_3^2)^{3/2} \le b_4 x_1 x_3 (x_1^2 + x_2^2).$$

From the above discussion, the truss design problem can be stated as the following nonlinear programming problem:

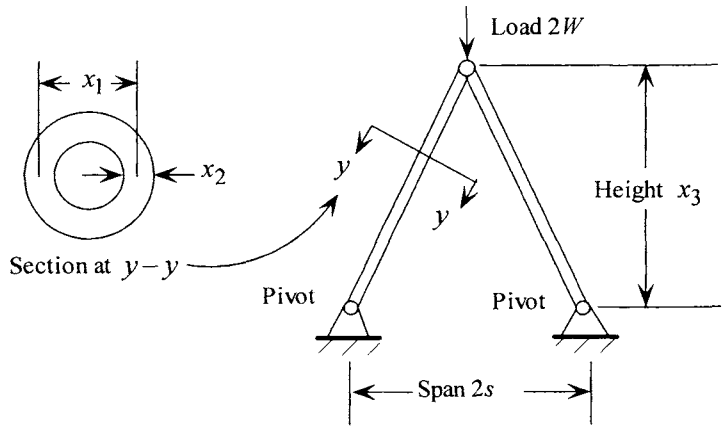


Figure 1.3 Two-bar truss.

$$
\begin{aligned}
\text{Minimize} \quad & x_1 x_2 (s^2 + x_3^2)^{1/2} \\
\text{subject to} \quad & x_3 - b_1 \le 0 \\
& x_1 - b_2 x_2 \le 0 \\
& W(s^2 + x_3^2)^{1/2} - b_3 x_1 x_2 x_3 \le 0 \\
& W(s^2 + x_3^2)^{3/2} - b_4 x_1 x_3 (x_1^2 + x_2^2) \le 0 \\
& x_1, x_2, x_3 \ge 0.
\end{aligned}
$$

C. Mechanical Design

In mechanical design, the concept of optimization can be used in conjunction with the traditional use of statics, dynamics, and the properties of materials. Asimov [1962], Fox [1971], and Johnson [1971] give several examples of optimal mechanical designs using mathematical programming. As noted by Johnson [1971], in designing mechanisms for high-speed machines, significant dynamic stresses and vibrations are inherently unavoidable. Hence, it is necessary to design certain mechanical elements on the basis of minimizing these undesirable characteristics. The following example illustrates an optimal design for a bearing journal.

Journal Design Problem Consider a two-bearing journal, each of length L, supporting a flywheel of weight W mounted on a shaft of diameter D, as shown in Figure 1.4. We wish to determine L and D that minimize frictional moment while keeping the shaft twist angle and clearances within acceptable limits.

A layer of oil film between the journal and the shaft is maintained by forced lubrication. The oil film serves to minimize the frictional moment and to

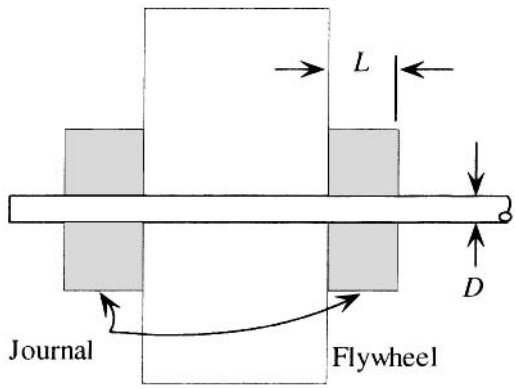


Figure 1.4 Journal bearing assembly.

limit the heat rise, thereby increasing the life of the bearing. Let h_0 be the smallest oil film thickness under steady-state operation. Then we must have

$$\hat{h}_0 \le h_0 \le \delta,$$

where h_0 is the minimum oil film thickness to prevent metal-to-metal contact and δ is the radial clearance specified as the difference between the journal radius and the shaft radius. A further limitation on h_0 is imposed by the following inequality:

$$0 \le e \le \hat{e},$$

where e is the *eccentricity ratio*, defined by $e = 1 - (h_0/\delta)$, and $\hat{e}$ is a prespecified upper limit.

Depending on the point at which the torque is applied on the shaft, or the nature of the torque impulses, and on the ratio of the shear modulus of elasticity to the maximum shear stress, a constant k_1 can be specified such that the angle of twist of the shaft is given by

$$\theta = \frac{1}{k_1 D}.$$

Furthermore, the frictional moment for the two bearings is given by

$$M = k_2 \frac{\omega}{\delta\sqrt{1-e^2}} D^3 L,$$

where k_2 is a constant that depends on the viscosity of the lubricating oil and ω is the rotational speed. Also, based on hydrodynamic considerations, the safe load-carrying capacity of a bearing is given by

$$c = k_3 \frac{\omega}{\delta^2} DL^3 \phi(e),$$

where k_3 is a constant depending on the viscosity of the oil and

$$\phi(e) = \frac{e}{(1-e^2)^2}[\pi^2(1-e^2) + 16e^2]^{1/2}.$$

Obviously, we need to have $2c \ge W$ to carry the weight W of the flywheel.

Thus, if δ, $\hat{h}_0$, and $\hat{e}$ are specified, one typical design problem is to find D, L, and h_0 to minimize the frictional moment while keeping the twist angle within an acceptable limit α. The model is thus given by:

$$\text{Minimize} \quad \frac{\omega}{\delta\sqrt{1-e^2}} D^3 L$$

$$\text{subject to} \quad \frac{1}{k_1 D} \le \alpha$$

$$2\frac{k_3\omega}{\delta^2} DL^3\phi\left(1-\frac{h_0}{\delta}\right) \ge W$$

$$\hat{h}_0 \le h_0 \le \delta$$

$$0 \le 1-\frac{h_0}{\delta} \le \hat{e}$$

$$D \ge 0$$

$$L \ge 0.$$

For a thorough discussion of this problem, the reader may refer to Asimov [1962]. The reader can also formulate the model to minimize the twist angle subject to the frictional moment being within a given maximum limit M'. We could also conceive of an objective function involving both the frictional moment and the angle of twist, if proper weights for these factors are selected to reflect their relative importance.

D. Electrical Networks

It has been well recognized for over a century that the equilibrium conditions of an electrical or a hydraulic network are attained as the total energy loss is minimized. Dennis [1959] was perhaps the first to investigate the relationship between electrical circuit theory, mathematical programming, and duality. The following discussion is based on his pioneering work.

An electrical circuit can be described by, for example, *n branches* connecting *m nodes.* In the following, we consider a direct-current network and assume that the nodes and each connecting branch are defined so that only one of the following electrical devices is encountered:

1. A *voltage source* that maintains a constant branch voltage v_s irrespective of the branch current c_s. Such a device absorbs power equal to $-v_s c_s$.

2. A *diode* that permits the branch current c_d to flow in only one direction and consumes zero power regardless of the branch current or voltage. Denoting the latter by v_d, this can be stated as

$$c_d \ge 0, \qquad v_d \ge 0, \qquad v_d c_d = 0. \tag{1.1}$$

3. A *resistor* that consumes power and whose branch current c_r and branch voltage v_r are related by

$$v_r = -rc_r, \tag{1.2}$$

where r is the *resistance* of the resistor. The power consumed is given by

$$-v_r c_r = \frac{v_r^2}{r} = rc_r^2. \tag{1.3}$$

The three devices are shown schematically in Figure 1.5. The current flow in the diagram is shown from the negative terminal of the branch to the positive terminal of the branch. The former is called the *origin node*, and the latter is the *ending node* of the branch. If the current flows in the opposite direction, the corresponding branch current will have a negative value, which, incidentally, is not permissible for the diode. The same sign convention will be used for branch voltages.

A network having a number of branches can be described by a *node-branch incidence matrix* **N**, whose rows correspond to the nodes and whose columns correspond to the branches. A typical element n_{ij} of **N** is given by

$$n_{ij} = \begin{cases} -1 & \text{if branch } j \text{ has node } i \text{ as its origin} \\ 1 & \text{if branch } j \text{ ends in node } i \\ 0 & \text{otherwise.} \end{cases}$$

For a network having several voltage sources, diodes, and resistors, let $\mathbf{N}_S$ denote the node-branch incidence matrix for all the branches having voltage sources, $\mathbf{N}_D$ denote the node-branch incidence matrix for all branches having diodes, and $\mathbf{N}_R$ denote the node-branch incidence matrix for all branches having resistors. Then, without loss of generality, we can partition **N** as

$$\mathbf{N} = [\mathbf{N}_S, \mathbf{N}_D, \mathbf{N}_R].$$

Similarly, the column vector **c**, representing the branch currents, can be partitioned as

$$\mathbf{c}^t = [\mathbf{c}_S^t, \mathbf{c}_D^t, \mathbf{c}_R^t],$$

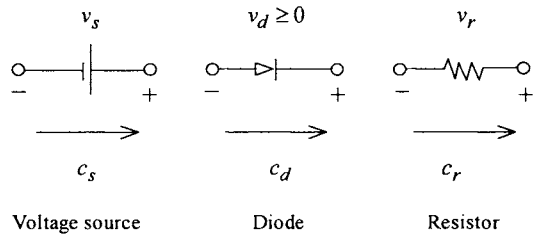


Figure 1.5 Typical electrical devices in a circuit.

and the column vector $\mathbf{v}$, representing the branch voltages, can be written as

$$\mathbf{v}^t = [\mathbf{v}_S^t, \mathbf{v}_D^t, \mathbf{v}_R^t].$$

Associated with each node i is a *node potential* p_i. The column vector $\mathbf{p}$, representing node potentials, can be written as

$$\mathbf{p}^t = [\mathbf{p}_S^t, \mathbf{p}_D^t, \mathbf{p}_R^t].$$

The following basic laws govern the equilibrium conditions of the network:

Kirchhoff's node law. The sum of all currents entering a node is equal to the sum of all currents leaving the node. This can be written as $\mathbf{Nc} = \mathbf{0}$, or

$$\mathbf{N}_S\mathbf{c}_S + \mathbf{N}_D\mathbf{c}_D + \mathbf{N}_R\mathbf{c}_R = \mathbf{0}. \tag{1.4}$$

Kirchhoff's loop law. The difference between the node potentials at the ends of each branch is equal to the branch voltage. This can be written as $\mathbf{N}^t\mathbf{p} = \mathbf{v}$, or

$$\begin{aligned} \mathbf{N}_S^t\mathbf{p} &= \mathbf{v}_S \\ \mathbf{N}_D^t\mathbf{p} &= \mathbf{v}_D \\ \mathbf{N}_R^t\mathbf{p} &= \mathbf{v}_R. \end{aligned} \tag{1.5}$$

In addition, we have the equations representing the characteristics of the electrical devices. From (1.1), for the set of diodes, we have

$$\mathbf{v}_D \geq \mathbf{0}, \quad \mathbf{c}_D \geq \mathbf{0}, \quad \mathbf{v}_D^t\mathbf{c}_D = 0, \tag{1.6}$$

and from (1.2), for the resistors, we have

$$\mathbf{v}_R = -\mathbf{R}\mathbf{c}_R, \tag{1.7}$$

where $\mathbf{R}$ is a diagonal matrix whose diagonal elements are the resistance values.

Thus, (1.4) – (1.7) represent the equilibrium conditions of the circuit, and we wish to find $\mathbf{v}_D$, $\mathbf{v}_R$, $\mathbf{c}$, and $\mathbf{p}$ satisfying these conditions.

Now, consider the following *quadratic programming problem*, which is discussed in Section 11.2:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\mathbf{c}_R^t\mathbf{R}\mathbf{c}_R - \mathbf{v}_S^t\mathbf{c}_S \\ \text{subject to} \quad & \mathbf{N}_S\mathbf{c}_S + \mathbf{N}_D\mathbf{c}_D + \mathbf{N}_R\mathbf{c}_R = \mathbf{0} \\ & -\mathbf{c}_D \leq \mathbf{0}. \end{aligned}$$

Here we wish to determine the branch currents $\mathbf{c}_S$, $\mathbf{c}_D$, and $\mathbf{c}_R$ to minimize the sum of half the energy absorbed in the resistors and the energy loss of the voltage source. From Section 4.3 the optimality conditions for this problem are

$$\begin{aligned} \mathbf{N}_S^t\mathbf{u} - \mathbf{v}_S &= \mathbf{0} \\ \mathbf{N}_D^t\mathbf{u} - \mathbf{I}\mathbf{u}_0 &= \mathbf{0} \\ \mathbf{N}_R^t\mathbf{u} + \mathbf{R}\mathbf{c}_R &= \mathbf{0} \\ \mathbf{N}_S\mathbf{c}_S + \mathbf{N}_D\mathbf{c}_D + \mathbf{N}_R\mathbf{c}_R &= \mathbf{0} \\ \mathbf{c}_D^t\mathbf{u}_0 &= 0 \\ \mathbf{c}_D, \mathbf{u}_0 &\geq \mathbf{0}, \end{aligned}$$

where $\mathbf{u}$ and $\mathbf{u}_0$ are column vectors representing the *Lagrangian multipliers*. It can readily be verified that letting $\mathbf{v}_D = \mathbf{u}_0$, $\mathbf{p} = \mathbf{u}$, and noting (1.7), the conditions above are precisely the equilibrium conditions (1.4) – (1.7). Note that the Lagrangian multiplier vector $\mathbf{u}$ is precisely the node potential vector $\mathbf{p}$.

Associated with the above problem is another problem, referred to as the *dual problem* (given below), where $\mathbf{G} = \mathbf{R}^{-1}$ is a diagonal matrix whose elements are the conductances and where $\mathbf{v}_S$ is fixed.

$$\begin{aligned} \text{Maximize } & -\frac{1}{2}\mathbf{v}_R^t\mathbf{G}\mathbf{v}_R \\ \text{subject to } & \mathbf{N}_S^t\mathbf{p} = \mathbf{v}_S \\ & \mathbf{N}_D^t\mathbf{p} - \mathbf{v}_D = \mathbf{0} \\ & \mathbf{N}_R^t\mathbf{p} - \mathbf{v}_R = \mathbf{0} \\ & \mathbf{v}_D \geq \mathbf{0}. \end{aligned}$$

Here, $\mathbf{v}_R^t\mathbf{G}\mathbf{v}_R$ is the power absorbed by the resistors, and we wish to find the branch voltages $\mathbf{v}_D$ and $\mathbf{v}_R$ and the potential vector $\mathbf{p}$.

The optimality conditions for this problem also are precisely (1.4)–(1.7). Furthermore, the Lagrangian multipliers for this problem are the branch currents.

It is interesting to note by Theorem 6.2.4, the main Lagrangian duality theorem, that the objective function values of the above two problems are equal at optimality; that is,

$$\frac{1}{2}\mathbf{c}_R^t\mathbf{R}\mathbf{c}_R + \frac{1}{2}\mathbf{v}_R^t\mathbf{G}\mathbf{v}_R - \mathbf{v}_S^t\mathbf{c}_S = 0.$$

Since $\mathbf{G} = \mathbf{R}^{-1}$ and noting (1.6) and (1.7), the above equation reduces to

$$\mathbf{v}_R^t\mathbf{c}_R + \mathbf{v}_D^t\mathbf{c}_D + \mathbf{v}_S^t\mathbf{c}_S = 0,$$

which is precisely the principle of energy conservation.

The reader may be interested in other applications of mathematical programming for solving problems associated with generation and distribution of electrical power. A brief discussion, along with suitable references, is given in the Notes and References section at the end of the chapter.

E. Water Resources Management

We now develop an optimization model for the conjunctive use of water resources for both hydropower generation and agricultural use. Consider the river basin depicted schematically in Figure 1.6.

A dam across the river provides the surface water storage facility to provide water for power generation and agriculture. The power plant is assumed to be close to the dam, and water for agriculture is conveyed from the dam, directly or after power generation, through a canal.

There are two classes of variables associated with the problem:

1. ***Design variables***: What should be the optimal capacity S of the reservoir, the capacity U of the canal supplying agricultural water, and the capacity E of the power plant?
2. ***Operational variables***: How much water should be released for agricultural power generation and for other purposes?

From Figure 1.6, the following operational variables can readily be identified for the jth period:

x_j^A = water released from the dam for agriculture

x_j^{PA} = water released for power generation and then for agricultural use

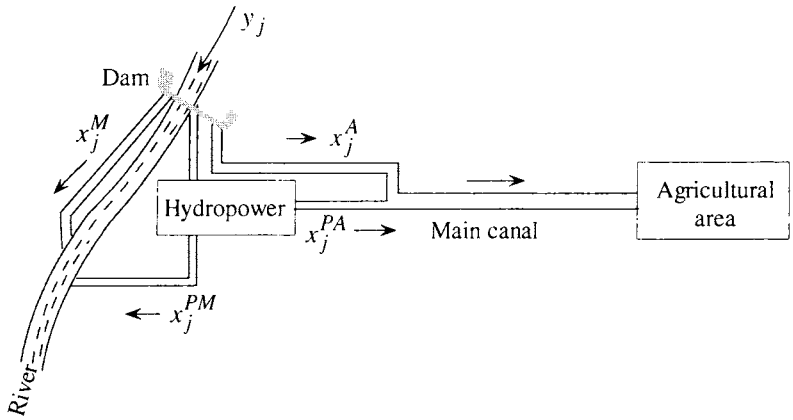


Figure 1.6 Typical river basin.

x_j^{PM} = water released for power generation and then returned downstream

x_j^{M} = water released from the dam directly downstream.

For the purpose of a planning model, we shall adopt a planning horizon of N periods, corresponding to the life span of major capital investments, such as that for the dam. The objective is to minimize the total discounted costs associated with the reservoir, power plant, and canal, minus the revenues from power generation and agriculture. These costs and revenues are discussed below.

Power Plant: Associated with the power plant, we have a cost of

$$C(E)+\sum_{j=1}^{N}\beta_j\hat{C}_e(E), \tag{1.8}$$

where $C(E)$ is the cost of the power plant, associated structures, and transmission facilities if the power plant capacity is E, and $\hat{C}_e(E)$ is the annual operation, maintenance, and replacement costs of the power facilities. Here, β_j is a discount factor that gives the present worth of the cost in period j. See Mobasheri [1968] for the nature of the functions $C(E)$ and $\hat{C}_e(E)$.

Furthermore, the discounted revenues associated with the energy sales can be expressed as

$$\delta\left\{\sum_{j=1}^{N}\beta_j[p_f F_j+p_d(f_j-F_j)]\right\}+(1-\delta)\left\{\sum_{j=1}^{n}\beta_j[p_f f_j-p_s(F_j-f_j)]\right\}, \tag{1.9}$$

where F_j is the known firm power demand that can be sold at p_f and f_j is the power production. Here $\delta = 1$ if $f_j > F_j$, and the excess power $f_j - F_j$ can be sold at a dump price of p_d. On the other hand, $\delta = 0$ if $f_j < F_j$, and a penalty of $p_s(F_j - f_j)$ is incurred since power has to be bought from adjoining power networks.

Reservoir and Canal: The discounted capital costs are given by

$$C_r(S)+\alpha C_\ell(U), \tag{1.10}$$

where $C_r(S)$ is the cost of the reservoir if its capacity is S, and $C_\ell(U)$ is the capital cost of the main canal if its capacity is U. Here α is a scalar to account for the lower life span of the canal compared to that of the reservoir.

The discounted operational costs are given by

$$\sum_{j=1}^{N}\beta_j[\hat{C}_r(S)+\hat{C}_\ell(U)]. \tag{1.11}$$

The interested reader may refer to Maass et al. [1967] and Mobasheri [1968] for a discussion on the nature of the functions discussed here.

Irrigation Revenues: The crop yield from irrigation can be expressed as a function R of the water used for irrigation during period j as shown by Minhas, et al. [1974]. Thus, the revenue from agriculture is given by

$$\sum_{j=1}^{N} \beta_j R(x_j^A + x_j^{PA}). \tag{1.12}$$

Here, for convenience, we have neglected the water supplied through rainfall.

Thus far we have discussed the various terms in the objective function. The model must also consider the constraints imposed on the design and decision variables.

Power Generation Constraints: Clearly, the power generated cannot exceed the energy potential of the water supplied, so that

$$f_j \le (x_j^{PM} + x_j^{PA})\Psi(s_j)\gamma e, \tag{1.13}$$

where $\Psi(s_j)$ is the head created by the water s_j stored in the reservoir during period j, γ is the power conversion factor, and e is the efficiency of the power system. (Refer to O'Laoghaine and Himmelblau [1974] for the nature of the function Ψ.)

Similarly, the power generated cannot exceed the generating capacity of the plant, so that

$$f_j \le \alpha_j E e H_j, \tag{1.14}$$

where α_j is the load factor defined as the ratio of the average daily production to the daily peak production and H_j is the number of operational hours.

Finally, the capacity of the plant has to be within known acceptable limits; that is,

$$E' \le E \le E''. \tag{1.15}$$

Reservoir Constraints: If we neglect the evaporation losses, the amount of water y_j flowing into the dam must be equal to the change in the amount stored in the dam and the water released for different purposes. This can be expressed as

$$s_{j+1} - s_j + x_j^A + x_j^M + x_j^{PM} + x_j^{PA} = y_j. \tag{1.16}$$

A second set of constraints states that the storage of the reservoir should be adequate and be within acceptable limits; that is,

$$S \ge s_j \tag{1.17}$$

$$S' \le S \le S''. \tag{1.18}$$

Mandatory Water Release Constraint: It is usually necessary to specify that a certain amount of water M_j is released to meet the downstream water requirements. This mandatory release requirement may be specified as

$$x_j^M + x_j^{PM} \ge M_j. \tag{1.19}$$

Canal Capacity: Finally, we need to specify that the canal capacity U should be adequate to handle the agricultural water. Hence,

$$x_j^A + x_j^{PA} \le U. \tag{1.20}$$

The objective is, then, to minimize the net costs represented by the sum of (1.8), (1.10), and (1.11), minus the revenues given by (1.9) and (1.12). The constraints are given by (1.13) to (1.20), together with the restriction that all variables are nonnegative.

F. Stochastic Resource Allocation

Consider the following linear programming problem:

$$\begin{aligned} &\text{Maximize } \mathbf{c}^t\mathbf{x} \\ &\text{subject to } \mathbf{A}\mathbf{x} \le \mathbf{b} \\ &\qquad\qquad \mathbf{x} \ge \mathbf{0}, \end{aligned}$$

where $\mathbf{c}$ and $\mathbf{x}$ are n-vectors, $\mathbf{b}$ is an m-vector, and $\mathbf{A} = [\mathbf{a}_1,...,\mathbf{a}_n]$ is an $m \times n$ matrix. The above problem can be interpreted as a resource allocation model as follows. Suppose that we have m resources represented by the vector $\mathbf{b}$. Column $\mathbf{a}_j$ of $\mathbf{A}$ represents an activity j, and the variable x_j represents the level of the activity to be selected. Activity j at level x_j consumes $\mathbf{a}_j x_j$ of the available resources; hence the constraint, $\mathbf{A}\mathbf{x} = \sum_{j=1}^n \mathbf{a}_j x_j \le \mathbf{b}$. If the unit profit of activity j is c_j, the total profit is $\sum_{j=1}^n c_j x_j = \mathbf{c}^t\mathbf{x}$. Thus, the problem can be interpreted as finding the best way of allocating the resource vector $\mathbf{b}$ to the various available activities so that the total profit is maximized.

For some practical problems, the above deterministic model is not adequate because the profit coefficients $c_1,...,c_n$ are not fixed but are random variables. We shall thus assume that $\mathbf{c}$ is a random vector with *mean* $\bar{\mathbf{c}} = (\bar{c}_1,..., \bar{c}_n)^t$ and *covariance matrix* $\mathbf{V}$. The objective function, denoted by z, will thus be a random variable with mean $\bar{\mathbf{c}}^t\mathbf{x}$ and variance $\mathbf{x}^t\mathbf{V}\mathbf{x}$.

If we want to maximize the expected value of z, we must solve the following problem:

$$\begin{aligned} &\text{Maximize } \bar{\mathbf{c}}^t\mathbf{x} \\ &\text{subject to } \mathbf{Ax} \leq \mathbf{b} \\ &\qquad\qquad \mathbf{x} \geq \mathbf{0}, \end{aligned}$$

which is a linear programming problem discussed in Section 2.6. On the other hand, if the variance of z is to be minimized, we have to solve the problem

$$\begin{aligned} &\text{Minimize } \mathbf{x}^t\mathbf{V}\mathbf{x} \\ &\text{subject to } \mathbf{Ax} \leq \mathbf{b} \\ &\qquad\qquad \mathbf{x} \geq \mathbf{0}, \end{aligned}$$

which is a *quadratic program* as discussed in Section 11.2.

Satisficing Criteria and Chance Constraints

In maximizing the expected value, we have completely neglected the variance of the gain z. On the other hand, while minimizing the variance, we did not take into account the expected value of z. In a realistic problem, one would perhaps like to maximize the expected value and, at the same time, minimize the variance. This is a multiple objective problem, and considerable research has been done on dealing with such problems (see Ehrgott [2004], Steur [1986], Zeleny [1974], and Zeleny and Cochrane [1973]). However, there are several other ways of considering the expected value and the variance simultaneously.

Suppose one is interested in ensuring that the expected value should be at least equal to a certain value $\bar{z}$, frequently referred to as an *aspiration level*, or a *satisficing level.* The problem can then be stated as:

$$\begin{aligned} &\text{Minimize } \mathbf{x}^t\mathbf{V}\mathbf{x} \\ &\text{subject to } \mathbf{Ax} \leq \mathbf{b} \\ &\qquad\qquad \bar{\mathbf{c}}^t\mathbf{x} \geq \bar{z} \\ &\qquad\qquad \mathbf{x} \geq \mathbf{0}, \end{aligned} \tag{1.21}$$

which is again a *quadratic programming problem.*

Another approach that can be adopted is as follows. Let $\alpha = \text{Prob}(\mathbf{c}^t\mathbf{x} \geq \bar{z})$; that is, α gives the probability that the aspiration level $\bar{z}$ will be attained. Clearly, one would like to maximize α. Now, suppose that the vector of random variables $\mathbf{c}$ can be expressed as the function $\mathbf{d} + y\mathbf{f}$, where $\mathbf{d}$ and $\mathbf{f}$ are fixed vectors and y is a random variable. Then

$$\begin{aligned} \alpha &= \text{Prob}(\mathbf{d}^t\mathbf{x} + y\mathbf{f}^t\mathbf{x} \geq \bar{z}) \\ &= \text{Prob}\left(y \geq \frac{\bar{z} - \mathbf{d}^t\mathbf{x}}{\mathbf{f}^t\mathbf{x}} \right) \end{aligned}$$

if $\mathbf{f}^t\mathbf{x} > 0$. Hence, in this case, the problem of maximizing α reduces to:

$$\begin{aligned} \text{Minimize} \quad & \frac{\bar{z} - \mathbf{d}^t\mathbf{x}}{\mathbf{f}^t\mathbf{x}} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \le \mathbf{b} \\ & \mathbf{x} \ge \mathbf{0}. \end{aligned}$$

This is a linear *fractional programming problem*, a solution procedure for which is discussed in Section 11.4.

Alternatively, if we wished to minimize the variance but we also wanted to include a constraint that required the probability of the profit $\mathbf{c}^t\mathbf{x}$ exceeding the desired value $\bar{z}$ to be at least some specified value q, this could be incorporated by using the following *chance constraint*:

$$\text{Prob}(\mathbf{c}^t\mathbf{x} \ge \bar{z}) = \text{Prob}\left(y \ge \frac{\bar{z} - \mathbf{d}^t\mathbf{x}}{\mathbf{f}^t\mathbf{x}} \right) \ge q.$$

Now assuming that y is a continuously distributed random variable for which ϕ_q denotes the upper $100q$ percentile value, that is, $\text{Prob}(y \ge \phi_q) = q$, the foregoing constraint can be written equivalently as

$$\frac{\bar{z} - \mathbf{d}^t\mathbf{x}}{\mathbf{f}^t\mathbf{x}} \le \phi_q \qquad \text{or} \qquad \mathbf{d}^t\mathbf{x} + \phi_q\mathbf{f}^t\mathbf{x} \ge \bar{z}.$$

This linear constraint can then be used to replace the expected value constraint in the model (1.21).

Risk Aversion Model

The approaches described above for handling the variance and the expected value of the return do not take into account the risk aversion behavior of individuals. For example, a person who wants to avoid risk may prefer a gain with an expected value of \$100 and a variance of 10 to a gain with an expected value of \$110 with variance of 30. A person who chooses the expected value of \$100 is more averse to risk than a person who might choose the alternative with an expected value of \$110. This difference in risk-taking behavior can be taken into account by considering the utility of money for the person.

For most people the value of an additional dollar decreases as their total net worth increases. The value associated with a net worth z is called the *utility* of z. Frequently, it is convenient to normalize the utility u so that $u = 0$ for $z = 0$ and $u = 1$ as z approaches the value ∞. The function u is called the person's utility function and is usually a nondecreasing continuous function. Figure 1.7 gives two typical utility functions for two people. For person (a), a gain of Δz

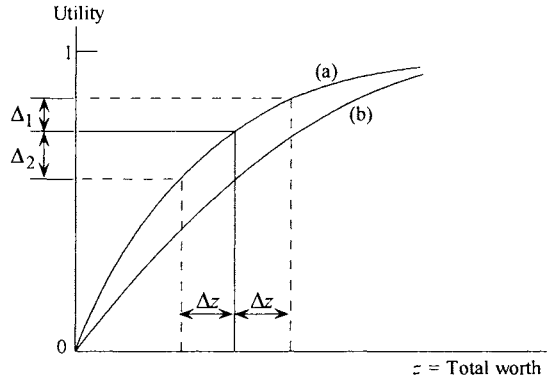


Figure 1.7 Utility functions.

increases the utility by Δ_1, and a loss of Δz decreases the utility by Δ_2. Since Δ_2 is larger than Δ_1, this person would prefer a lower variance. Such a person is more averse to risk than a person whose utility function is as in (b) in Figure 1.7.

Different curves, such as (a) or (b) in Figure 1.7, can be expressed mathematically as

$$u(z) = 1 - e^{-kz},$$

where $k > 0$ is called a *risk aversion constant.* Note that a larger value of k results in a more risk-averse behavior.

Now suppose that the current worth is zero, so that the total worth is equal to the gain z. Suppose that $\mathbf{c}$ is a normal random vector with mean $\bar{\mathbf{c}}$ and covariance matrix $\mathbf{V}$. Then z is a normal random variable with mean $\bar{z} = \bar{\mathbf{c}}^t\mathbf{x}$ and variance $\sigma^2 = \mathbf{x}^t\mathbf{V}\mathbf{x}$. In particular, the density function ϕ of the gain is given by

$$\phi(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-\bar{z}}{\sigma}\right)^2\right].$$

We wish to maximize the expected value of the utility given by

$$\int_{-\infty}^{\infty}\left(1-e^{-kz}\right)\phi(z)\,dz$$

$$= 1 - \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[-kz - \frac{1}{2}\left(\frac{z-\bar{z}}{\sigma}\right)^2\right]dz$$

$$= 1 - \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left(\frac{z-\bar{z}+k\sigma^2}{\sigma}\right)^2\right]\exp\left(-k\bar{z} + \frac{1}{2}k^2\sigma^2\right)dz$$

$$= 1 - \frac{\exp\left(-k\bar{z} + \frac{1}{2}k^2\sigma^2\right)}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{z - \bar{z} + k\sigma^2}{\sigma}\right)^2\right] dz$$

$$= 1 - \exp\left(-k\bar{z} + \frac{1}{2}k^2\sigma^2\right).$$

Hence, maximizing the expected value of the utility is equivalent to maximizing $k\bar{z} - (1/2)k^2\sigma^2$. Substituting for $\bar{z}$ and σ^2, we get the following *quadratic program*:

$$\begin{aligned} \text{Maximize} \quad & k\bar{\mathbf{c}}^t\mathbf{x} - \frac{1}{2}k^2\mathbf{x}^t\mathbf{V}\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned}$$

Again, this can be solved by using the methods discussed in Chapter 11, depending on the nature of $\mathbf{V}$.

G. Location of Facilities

A frequently encountered problem is the optimal location of centers of activities. This may involve the location of machines or departments in a factory, the location of factories or warehouses from which goods can be shipped to retailers or consumers, or the location of emergency facilities (i.e., fire or police stations) in an urban area.

Consider the following simple case. Suppose that there are n markets with known demands and locations. These demands are to be met from m warehouses of known capacities. The problem is to determine the locations of the warehouses so that the total distance weighted by the shipment from the warehouses to the markets is minimized. More specifically, let

(x_i, y_i) = unknown location of warehouse i for $i = 1,\ldots, m$

c_i = capacity of warehouse i for $i = 1,\ldots, m$

(a_j, b_j) = known location of market j for $j = 1,\ldots, n$

r_j = known demand at market j for $j = 1,\ldots, n$

d_{ij} = distance from warehouse i to market area j for $i = 1,\ldots, m$; $j = 1,\ldots, n$

w_{ij} = units shipped from warehouse i to market area j for $i = 1,\ldots, m$; $j = 1,\ldots, n$

The problem of locating the warehouses and determining the shipping pattern can be stated as follows:

$$\text{Minimize} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}d_{ij}$$

$$\text{subject to} \quad \sum_{j=1}^{n} w_{ij} \le c_i \qquad \text{for } i = 1,\ldots,m$$

$$\sum_{i=1}^{m} w_{ij} = r_j \qquad \text{for } j = 1,\ldots,n$$

$$w_{ij} \ge 0 \qquad \text{for } i = 1,\ldots,m;\ j = 1,\ldots,n.$$

Note that both w_{ij} and d_{ij} are to be determined, and hence, the above problem is a nonlinear programming problem. Different measures of distance can be chosen, using the *rectilinear*, *Euclidean*, or ℓ_p *norm metrics* , where the value of p could be chosen to approximate particular city travel distances. These are given respectively by

$$d_{ij} = |x_i - a_j| + |y_i - b_j|$$

$$d_{ij} = \left[(x_i - a_j)^2 + (y_i - b_j)^2\right]^{1/2}$$

$$d_{ij} = \left[(x_i - a_j)^p + (y_i - b_j)^p\right]^{1/p}.$$

Each choice leads to a particular nonlinear problem in the variables $x_1,\ldots,x_m$, $y_1,\ldots,y_m$, $w_{11},\ldots,w_{mn}$. If the locations of the warehouses are fixed, the d_{ij} values are known, and the above problem reduces to a special case of a linear programming problem known as the *transportation problem.* On the other hand, for fixed values of the transportation variables, the problem reduces to a (pure) *location problem.* Consequently, the above problem is also known as a *location-allocation problem.*

H. Miscellaneous Applications

There are a host of other applications to which nonlinear programming models and techniques have been applied. These include the problems of chemical equilibrium and process control; gasoline blending; oil extraction, blending, and distribution; forest thinning and harvest scheduling; economic equilibration of supply and demand interactions under various market behavioral phenomena; pipe network design for reliable water distribution systems; electric utility capacity expansion planning and load management; production and inventory control in manufacturing concerns; least squares estimation of statistical parameters and data fitting; and the design of engines, aircraft, ships, bridges, and other structures. The Notes and References section cites several references that provide details on these and other applications.

1.3 Guidelines for Model Construction

The modeling process is concerned with the construction of a mathematical abstraction of a given problem that can be analyzed to produce meaningful answers that guide the decisions to be implemented. Central to this process is the *identification* or the *formulation* of the problem. By the nature of human activities, a problem is seldom isolated and crisply defined, but rather, interacts with various other problems at the fringes and encompasses various details obfuscated by uncertainty. For example, a problem of scheduling jobs on machines interacts with the problems of acquiring raw materials, forecasting uncertain demand, and planning for inventory storage and dissipation; and it must contend with machine reliability, worker performance and absenteeism, and insertions of spurious or rush-jobs. A modeler must therefore identify the particular scope and aspect of the problem to be explicitly considered in formulating the problem, and must make suitable simplifying assumptions so that the resulting model is a balanced compromise between *representability* and *mathematical tractability*. The model, being only an abstraction of the real problem, will yield answers that are only as meaningful as the degree of accuracy with which it represents the actual physical system. On the other hand, an unduly complicated model might be too complex to be analyzed mathematically for obtaining any credible solution for consideration at all! This compromise, of course, need not be achieved at a single attempt. Often, it is instructive to begin with a simpler model representation, to test it to gain insights into the problem, and then to guide the direction in which the model should be further refined to make it more representative while maintaining adequate tractability. While accomplishing this, it should be borne in mind that the answers from the model are meant to provide guidelines for making decisions rather than to replace the decision maker. The model is only an abstraction of reality and is not necessarily an equivalent representation of reality itself. At the same time, these guidelines need to be well founded and meaningful. Moreover, one important function of a model is to provide more information on system behavior through *sensitivity analyses*, in which the response of the system is studied under various scenarios related to perturbations in different problem parameters. To obtain reliable insights through such an analysis, it is important that a careful balance be struck between problem representation and tractability.

Accompanying the foregoing process is the actual *construction of a mathematical statement* of the problem. Often, there are several ways in which an identified problem can be modeled mathematically. Although these alternative forms may be mathematically equivalent, they might differ substantially in the felicity they afford to solution algorithms. Hence, some foresight into the operation and limitations of algorithms is necessary. For example, the restriction that a variable x should take on the values 0, 1, or 2 can be modeled "correctly" using the constraint $x(x-1)(x-2)=0$. However, the nonconvex structure of this constraint will impose far more difficulty for most algorithms (unless the algorithm is designed to exploit such a polynomial structure) than if this discrete restriction was handled separately and explicitly as in a branch-and-bound framework, for instance (see Nemhauser and Wolsey [1998] or Parker and

Rardin [1988]). As another example, a feasible region defined by the inequalities $g_i(\mathbf{x}) \leq 0$ for $i = 1,..., m$ can be stated equivalently as the set of equality constraints $g_i(\mathbf{x}) + s_i^2 = 0$ for $i = 1,..., m$ by introducing new (unrestricted) variables s_i, $i = 1,..., m$. Although this is sometimes done to extend a theory or technique for equality constraints to one for inequality constraints, blind application of this strategy can be disastrous for solution algorithms. Besides increasing the dimension with respect to nonlinearly appearing variables, this modeling approach injects nonconvexities into the problem by virtue of which the optimality conditions of Chapter 4 can be satisfied at nonoptimal points, even though this might not have been the case with the original inequality-constrained problem.

In the same spirit, the inequality and equality constraints of the nonlinear program stated in Section 1.1 can be written equivalently as the *single* equality constraint

$$\sum_{i=1}^{m} [g_i(\mathbf{x}) + s_i^2]^2 + \sum_{j=1}^{\ell} h_j^2(\mathbf{x}) = 0,$$

or as

$$\sum_{i=1}^{m} \max\{g_i(\mathbf{x}), 0\} + \sum_{j=1}^{\ell} |h_j(\mathbf{x})| = 0,$$

or

$$\sum_{i=1}^{m} \max^2\{g_i(\mathbf{x}), 0\} + \sum_{j=1}^{\ell} h_j^2(\mathbf{x}) = 0.$$

These different statements have different structural properties; and if they are not matched properly with algorithmic capabilities, one can obtain meaningless or arbitrary solutions, if any at all. However, although such an equivalent single constraint is rarely adopted in practice, the conceptual constructs of these reformulations are indeed very useful in devising penalty functions when such equivalent constraint expressions are accommodated within the objective function, as we shall see in Chapter 9. Also, this underscores the need for knowing the underlying theory of nonlinear programming in order to be able to apply it appropriately in practice and to interpret the outputs produced from software. In other words, *one needs to be a good theoretician in order to be a good practitioner*. Of course, the converse of this statement also has merit.

Generally speaking, there are some guidelines that one can follow to construct a suitable mathematical formulation that will be amenable to most algorithms. Some experience and forethought is necessary in applying these guidelines, and the process is more of an art than a science. We provide some suggestions below but caution the reader that these are only general recommendations and guiding principles rather than a universal set of instructions.

Foremost among these guidelines are the requirements to construct an adequate statement of the problem, to identify any inherent special structures, and to exploit these structures in the algorithmic process. Such structures might simply be the linearity of constraints or the presence of tight lower and upper bounds on the variables, dictated either by practice or by some knowledge of the neighborhood containing an optimum. Most existing powerful algorithms require differentiability of the functions involved, so a smooth representation with derivative information is useful wherever possible. Although higher, second-order, derivative information is usually expensive to obtain and might require excessive storage for use in relatively large problems, it can enhance algorithmic efficiency substantially if available. Hence, many efficient algorithms use approximations of this information, assuming second-order differentiability. Besides linearity and differentiability, there are many other structures afforded by either the nature of the constraints themselves (such as network flow constraints) or, generally, by the manner in which the nonzero coefficients appear in the constraints (e.g., in a block diagonal fashion over a substantial set of constraints; see Lasdon [1970]). Such structures can enhance algorithmic performance and therefore can increase the size of problems that are solvable within a reasonable amount of computational effort.

In contrast with special structures that are explicitly identified and exploited, the problem function being optimized might be a complex "black-box" of an implicit unknown form whose evaluation itself might be an expensive task, perhaps requiring experimentation. In such instances, a response surface fitting methodology as described in Myers [1976] or some discretized grid approximations of such functions might be useful devices.

Also, quite often in practice, the objective function can be relatively flat in the vicinity of an optimum. After determining the optimal objective values, the given objective function could be transferred to the set of constraints by requiring to take on near-optimal values, thereby providing the opportunity to reoptimize with respect to another secondary objective function. This concept can be extended to multiple objective functions. This approach is known as a *preemptive priority strategy* for considering a hierarchy of prioritized multiple objective functions.

In the modeling process it is also useful to distinguish between *hard constraints*, which must necessarily be satisfied without any compromise, and *soft constraints*, for which mild violations can be tolerated, albeit at some incurred cost. For example, the expenditure $g(\mathbf{x})$ for some activity vector $\mathbf{x}$ might be required to be no more than a budgeted amount B, but violations within limits might be permissible if economically justifiable. Hence, this constraint can be modeled as $g(\mathbf{x}) - B = y^+ - y^-$, where y^+ and y^- are nonnegative variables, and where the "violation" y^+ is bounded above by a limit on the capital that can be borrowed or raised and, accordingly, also accompanied by a cost term $c(y^+)$ in the objective function. Such constraints are also referred to as *elastic constraints* because of the flexibility they provide.

It is insightful to note that permitting mild violations in some constraints, if tolerable, can have a significant impact on the solution obtained. For example, imposing a pair of constraints $h_1(\mathbf{x}) = 0$ and $h_2(\mathbf{x}) = 0$ as hard constraints might cause the feasible region defined by their intersection to be far removed from attractively valued solutions, while such solutions only mildly violate these constraints. Hence, by treating them as soft constraints and rewriting them as $-\Delta_i \leq h_i(\mathbf{x}) \leq \Delta_i$, where Δ_i is a small positive tolerance factor for $i = 1, 2$, we might be able to obtain far better solutions which, from a managerial viewpoint, compromise more judiciously between solution quality and feasibility. These concepts are related to *goal programming* (see Ignizio [1976]), where the soft constraints represent goals to be attained, along with accompanying penalties or rewards for under- or over-achievements.

We conclude this section by addressing the all-important but often neglected practice of problem *bounding* and *scaling*, which can have a profound influence on algorithmic performance. Many algorithms for both continuous and discrete optimization problems often benefit greatly by the presence of tight lower and upper bounds on variables. Such bounds could be constructed based on practical, optimality-based, or feasibility-based considerations. In addition, the operation of scaling deserves close attention. This can involve both the scaling of constraints by multiplying through with a (positive) constant, and the scaling of variables through a simple linear transformation that replaces $\mathbf{x}$ by $\mathbf{y} = \mathbf{D}\mathbf{x}$, where $\mathbf{D}$ is a nonsingular diagonal matrix. The end result sought is to try to improve the structural properties of the objective function and constraints, and to make the magnitudes of the variables, and the magnitudes of the constraint coefficients (as they dictate the values of the dual variables or Lagrange multipliers; see Chapter 4), vary within similar or compatible ranges. This tends to reduce numerical accuracy problems and to alleviate ill-conditioning effects associated with severely skewed or highly ridge-like function contours encountered during the optimization process. As can well be imagined, if a pipe network design problem, for example, contains variables representing pipe thicknesses, pipe lengths, and rates of flows, all in diversely varying dimensional magnitudes, this can play havoc with numerical computations. Besides, many algorithms base their termination criteria on prespecified tolerances on constraint satisfaction and on objective value improvements obtained over a given number of most recent iterations. Evidently, for such checks to be reliable, it is necessary that the problem be reasonably well scaled. This is true even for *scale-invariant algorithms*, which are designed to produce the same sequence of iterates regardless of problem scaling, but for which similar feasibility and objective improvement termination tests are used. Overall, although a sufficiently badly scaled problem can undoubtedly benefit by problem scaling, the effect of the scaling mechanism used on reasonably well-scaled problems can be mixed. As pointed out by Lasdon and Beck [1981], the scaling of nonlinear programs is as yet a "black art" that needs further study and refinement.

Exercises

[1.1] Consider the following nonlinear programming problem:

$$\begin{aligned} \text{Minimize} \quad & (x_1 - 4)^2 + (x_2 - 2)^2 \\ \text{subject to} \quad & 4x_1^2 + 9x_2^2 \leq 36 \\ & x_1^2 + 4x_2 = 4 \\ & \mathbf{x} = (x_1, x_2) \in X \equiv \{\mathbf{x} : 2x_1 \geq -3\}. \end{aligned}$$

a. Sketch the feasible region and the contours of the objective function. Hence, identify the optimum graphically on your sketch.
b. Repeat part a by replacing minimization with maximization in the problem statement.

[1.2] Suppose that the daily demand for product j is d_j for $j = 1, 2$. The demand should be met from inventory, and the latter is replenished from production whenever the inventory reaches zero. Here, the production time is assumed to be insignificant. During each product run, Q_j units can be produced at a fixed setup cost of $\$k_j$ and a variable cost of $\$c_jQ_j$. Also, a variable inventory-holding cost of $\$h_j$ per unit per day is also incurred, based on the average inventory. Thus, the total cost associated with product j during T days is $\$Td_jk_j/Q_j + Tc_jd_j + TQ_jh_j/2$. Adequate storage area for handling the maximum inventory Q_j has to be reserved for each product j. Each unit of product j needs s_j square feet of storage space, and the total space available is S.

a. We wish to find optimal production quantities Q_1 and Q_2 to minimize the total cost. Construct a model for this problem.
b. Now suppose that shortages are permitted and that production need not start when inventory reaches a level of zero. During the period when inventory is zero, demand is not met and the sales are lost. The loss per unit thus incurred is $\$\ell_j$. On the other hand, if a sale is made, the profit per unit is $\$P_j$. Reformulate the mathematical model.

[1.3] A manufacturing firm produces four different products. One of the necessary raw materials is in short supply, and only R pounds are available. The selling price of product i is $\$S_i$ per pound. Furthermore, each pound of product i uses a_i pounds of the critical raw material. The variable cost, excluding the raw material cost, of producing x_i pounds of product i is $k_ix_i^2$, where $k_i > 0$ is known. Develop a mathematical model for the problem.

[1.4] Suppose that the demand $d_1,...,d_n$ for a certain product over n periods is known. The demand during period j can be met from the production x_j during the period or from the warehouse stock. Any excess production can be stored at the warehouse. However, the warehouse has capacity K, and it would cost $\$c$ to carry over one unit from one period to another. The cost of production during period j is given by $f(x_j)$ for $j = 1,..., n$. If the initial inventory is I_0, formulate the production scheduling problem as a nonlinear program.

[1.5] An office room of length 70 feet and width 45 feet is to be illuminated by n light bulbs of wattage W_i, $i = 1,..., n$. The bulbs are to be located 7 feet above the working surface. Let (x_i, y_i) denote the x and y coordinates of the ith bulb. To ensure adequate lighting, illumination is checked at the working surface level at grid points of the form (α, β), where

$$\alpha = 10p, \quad p = 0, 1,..., 7$$

$$\beta = 5q, \quad q = 0, 1,..., 9.$$

The illumination at (α, β) resulting from a bulb of wattage W_i located at (x_i, y_i) is given by

$$E_i(\alpha,\beta) = k\frac{W_i \left\|(\alpha,\beta) - (x_i, y_i)\right\|}{\left\|(\alpha,\beta,7) - (x_i, y_i, 0)\right\|^3},$$

where k is a constant reflecting the efficiency of the bulb. The total illumination at (α, β) can be taken to be $\sum_{i=1}^{n} E_i(\alpha,\beta)$. At each of the points checked, an illumination of between 3.2 and 5.6 units is required. The wattage of the bulbs used is between 60 and 300 W. Assume that $W_i \ \forall i$ are continuous variables.

a. Construct a mathematical model to minimize the number of bulbs used and to determine their location and wattage, assuming that the cost of installation and of periodic bulb replacement is a function of the number of bulbs used.
b. Construct a mathematical model similar to that of part a, with the added restriction that all bulbs must be of the same wattage.

[1.6] Consider the following *portfolio selection problem.* An investor must choose a portfolio $\mathbf{x} = (x_1, x_2,..., x_n)^t$, where x_j is the proportion of the assets allocated to the jth security. The return on the portfolio has mean $\bar{\mathbf{c}}^t\mathbf{x}$ and variance $\mathbf{x}^t\mathbf{V}\mathbf{x}$, where $\bar{\mathbf{c}}$ is the vector denoting mean returns and $\mathbf{V}$ is the matrix of covariances of the returns. The investor would like to increase his or her expected return while decreasing the variance and hence the risk. A portfolio is called *efficient* if there exists no other portfolio having a larger expected return and a smaller variance. Formulate the problem of finding an efficient portfolio, and suggest some procedures for choosing among efficient portfolios.

[1.7] A household with budget b purchases n commodities. The unit price of commodity j is c_j, and the minimal amount of the commodity to be purchased is ℓ_j. After the minimal amounts of the n products are consumed, a function a_j of the remaining budget is allocated to commodity j. The behavior of the household is observed over m months for the purpose of estimating $\ell_1,...,\ell_n$, and $a_1,...,a_n$. Develop a regression model for estimating these parameters if:

a. The sum of the squares of the error is to be minimized.
b. The maximum absolute value of the error is to be minimized.
c. The sum of the absolute values of the error is to be minimized.
d. For both parts b and c, reformulate the problems as linear programs.

[1.8] A rectangular heat storage unit of length L, width W, and height H will be used to store heat energy temporarily. The rate of heat losses h_c due to convection and h_r due to radiation are given by

$$h_c = k_c A(T - T_a)$$

$$h_r = k_r A(T^4 - T_a^4),$$

where k_c and k_r are constants, T is the temperature of the heat storage unit, A is the surface area, and T_a is the ambient temperature. The heat energy stored in the unit is given by

$$Q = kV(T - T_a),$$

where k is a constant and V is the volume of the storage unit. The storage unit should have the ability to store at least Q'. Furthermore, suppose that space availability restricts the dimensions of the storage unit to

$$0 \le L \le L', \qquad 0 \le W \le W', \qquad \text{and} \qquad 0 \le H \le H'.$$

a. Formulate the problem of finding the dimensions L, W, and H to minimize the total heat losses.
b. Suppose that the constants k_c and k_r are linear functions of t, the insulation thickness. Formulate the problem of determining optimal dimensions L, W, and H to minimize the insulation cost.

[1.9] Formulate the model for Exercise 1.8 if the storage unit is a cylinder of diameter D and height H.

[1.10] Suppose that the demand for a certain product is a normally distributed random variable with mean 150 and variance 49, and that the production function is given by $p(\mathbf{x}) = \boldsymbol{\alpha}^t \mathbf{x}$, where $\mathbf{x}$ represents a set of n activity levels. Formulate the chance constraint that the probability of production falling short of demand by more than 5 units should be no more than 1% as a linear constraint.

[1.11] Consider a linear program to minimize $\mathbf{c}^t\mathbf{x}$ subject to $\mathbf{Ax} \leq \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$. Suppose that the components c_j of the vector $\mathbf{c}$ are random variables distributed independently of each other and of the $\mathbf{x}$-variables, and that the expected value of c_j is $\bar{c}_j, j = 1,..., n$.

a. Show that the minimum expected cost is obtained by solving the problem to minimize $\bar{\mathbf{c}}^t\mathbf{x}$ subject to $\mathbf{Ax} \leq \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$, where $\bar{\mathbf{c}} = (\bar{c}_1,...,\bar{c}_n)^t$.

b. Suppose that a firm makes two products that consume a common resource, which is expressed as follows:

$$5x_1 + 6x_2 \leq 30,$$

where x_j is the amount of product j produced. The unit profit for product 1 is normally distributed with mean 4 and variance 2. The unit profit for product 2 is given by a χ^2-distribution with 2 degrees of freedom. Assume that the random variables are independently distributed and that they are not dependent upon x_1 and x_2. Find the quantities of each product that must be produced to maximize expected profit. Will your answer differ if the variance for the first product were 4?

[1.12] Consider the following problem of a regional effluent control along a river. Currently, n manufacturing facilities dump their refuse into the river. The current rate of dumping by facility j is $\mu_j, j = 1,..., n$. The water quality is examined along the river at m control points. The minimum desired quality improvement at point i is $b_i, i = 1,..., m$. Let x_j be the amount of waste to be removed from source j at a cost of $f_j(x_j)$, and let a_{ij} be the quality improvement at control point i for each unit of waste removed at source j.

a. Formulate the problem of improving the water quality at a minimum cost as a nonlinear program.

b. In the above formulation, it is possible that certain sources would have to remove substantial amounts of waste, whereas others would only be required to remove small amounts of waste or none at all. Reformulate the problem so that a measure of equity among the sources is attained.

[1.13] A steel company manufactures crankshafts. Previous research indicates that the mean shaft diameter may assume the value μ_1 or μ_2, where $\mu_2 > \mu_1$. Furthermore, the probability that the mean is equal to μ_1 is p. To test whether the mean is μ_1 or μ_2, a sample of size n is chosen, and the diameters $x_1,...,x_n$ are recorded. If $\bar{x} = \sum_{j=1}^{n} x_j/n$ is less than or equal to K, the hypothesis $\mu = \mu_1$ is accepted; otherwise, the hypothesis $\mu = \mu_2$ is accepted. Let $f(\bar{x} \mid \mu_1)$ and $f(\bar{x} \mid \mu_2)$ be the probability density functions of the sample mean if the popula-

tion mean is μ_1 and μ_2, respectively. Furthermore, suppose that the penalty cost of accepting $\mu = \mu_1$ when $\mu = \mu_2$ is α and that the penalty cost of accepting $\mu = \mu_2$ when $\mu = \mu_1$ is β. Formulate the problem of choosing K such that the expected total cost is minimized. Show how the problem could be reformulated as a nonlinear program.

[1.14] An elevator has a vertical acceleration $u(t)$ at time t. Passengers would like to move from the ground level at zero altitude to the sixteenth floor at altitude 50 as fast as possible but dislike fast acceleration. Suppose that the passenger's time is valued at $\$\alpha$ per unit time, and furthermore, suppose that the passenger is willing to pay at a rate of $\$\beta u^2(t)$ per unit time to avoid fast acceleration. Formulate the problem of determining the acceleration from the time the elevator starts ascending until it reaches the sixteenth floor as an optimal control problem. Can you formulate the problem as a nonlinear program?

Notes and References

The advent of high-speed computers has considerably increased our ability to apply iterative procedures for solving large-scale problems, both linear and nonlinear. Although our ability to obtain global minimal solutions to nonconvex problems of realistic size is still rather limited, continued theoretical breakthroughs are overcoming this handicap (see Horst and Tuy [1993], Horst et al. [2000], Sherali and Adams [1999], and Zabinski [2003].)

Section 1.2 gives some simplified examples of problems that could be solved by the nonlinear programming methods discussed in the book. Our purpose was not to give complete details but only a flavor of the diverse problem areas that can be attacked. See Lasdon and Waren [1980] for further applications.

Optimal control is closely linked with mathematical programming. Dantzig [1966] has shown how certain optimal control problems can be solved by applying the simplex method. For further details of the application of mathematical programming to control problems, refer to Bracken and McCormick [1968], Canon and Eaton [1966], Canon et al. [1970], Cutler and Perry [1983], and Tabak and Kuo [1971].

With the recent developments and interest in aerospace and related technology, optimum design in this area has taken on added importance. In fact, since 1969, the Advisory Group for Aerospace Research and Development under NATO has sponsored several symposia on structural optimization. With improved materials being used for special purposes, optimum mechanical design has also increased in importance. The works of Cohn [1969], Fox [1969, 1971], Johnson [1971], Majid [1974], and Siddal [1972] are of interest in understanding how design concepts are integrated with optimization concepts in mechanical and structural design. Also, see Sherali and Ganesan [2003] (and the references cited therein) for ship design problems and related response surface methodological approaches.

Mathematical programming has also been used successfully to solve various problems associated with the generation and distribution of electrical

power and the operation of the system. These problems include the study of load flow, substation switching, expansion planning, maintenance scheduling, and the like. In the load flow problem, one is concerned with the flow of power through a transmission network to meet a given demand. The power distribution is governed by the well-known Kirchhoff's laws, and the equilibrium power flows satisfying these conditions can be computed by nonlinear programming. In other situations, the power output from hydroelectric plants is considered fixed, and the objective is to minimize the cost of fuel at the thermal plants. This problem, referred to as the economic dispatch problem, is usually solved online every few minutes, with appropriate power adjustments made. The generation capacity expansion problems study a minimum-cost equipment purchase and dispatchment plan that can satisfy the demand load at a specified reliability level over a given time horizon. For more details, refer to Abou-Taleb et al. [1974], Adams et al. [1972], Anderson [1972], Beglari and Laughton [1975], Bloom [1983], Bloom et al. [1984], Kirchmayer [1958], Sasson [1969a and 1969b], Sasson and Merrill [1974], Sasson et al. [1971], Sherali [1985], Sherali and Soyster [1983], and Sherali and Staschus [1985].

The field of water resources systems analysis has shown spectacular growth during the last three decades. As in many fields of science and technology, the rapid growth of water resources engineering and systems analysis was accompanied by an information explosion of considerable proportions. The problem discussed in Section 1.2 is concerned with rural water resources management for which an optimal balance between the use of water for hydropower generation and agriculture is sought. Some typical studies in this area can be found in Haimes [1973, 1977], Haimes and Nainis [1974], and Yu and Haimes [1974].

As a result of the rapid growth of urban areas, city managers are also concerned with integrating urban water distribution and land use. Some typical quantitative studies on urban water distribution and disposal may be found in Argaman et al. [1973], Dajani et al. [1972], Deb and Sarkar [1971], Fujiwara et al. [1987], Jacoby [1968], Loganathan et al. [1990], Shamir [1974], Sherali et al. [2001], Walsh and Brown [1973], and Wood and Charles [1973].

In his classic study on portfolio allocation, Markowitz [1952] showed how the variance of the returns on the portfolio can be incorporated in the optimal decision. In Exercise 1.6 the portfolio allocation problem is introduced briefly.

From 1955 to 1959, numerous studies were undertaken to incorporate uncertainty in the parameter values of a linear program. Refer to Charnes and Cooper [1959] , Dantzig [1955], Freund [1956], and Madansky [1959] for some of the early work in this area. Since then, many other studies have been conducted. The approaches, referred to in the literature as *chance constrained problems* and *programming with recourse*, seem particularly attractive. The interested reader may refer to Charnes and Cooper [1961, 1963], Charnes et al. [1967], Dantzig [1963], Elmaghraby [1960], Evers [1967], Geoffrion [1967c], Madansky [1962], Mangasarian [1964], Parikh [1970], Sengupta [1972], Sengupta and Portillo-Campbell [1970], Sengupta et al. [1963], Vajda [1970, 1972], Wets [1966a, 1966b, 1972], Williams [1965, 1966], and Ziemba [1970,

1971, 1974, 1975]. Also, see Mulvey et al. [1995] and Takriti and Ahmed [2004] for robust optimization models and Sen and Higle [2000] for stochastic optimization approaches.

For a description of other applications, the interested reader is referred to Ali et al. [1978] for an oil resource management problem; to Lasdon [1985] and Prince et al. [1983] for Texaco's OMEGA gasoline blending problem; to Rothfarb et al. [1970] for the design of offshore natural gas pipeline distribution systems; to Berna et al. [1980], Heyman [1990], Sarma and Reklaitis [1979], and Wall et al. [1986] for chemical process optimization and equilibrium problems; to Intriligator [1971], Murphy et al. [1982], Sherali [1984], and Sherali et al. [1983] for mathematical economics problems; to Adams and Sherali [1984], Francis et al. [1991], Love et al. [1988], Sherali and Tuncbilek [1992], Sherali et al. [2002], and Shetty and Sherali [1980] for location-allocation problems; to Bullard et al. [1985] for forest harvesting problems; to Jones [2001] and Myers [1976] for response surface methodologies; and to Dennis and Schnabel [1983], Fletcher [1987], and Sherali et al. [1988] for a discussion on least squares estimation problems with applications to data fitting and statistical parameter estimation.

For further discussion on problem scaling we refer the reader to Bauer [1963], Curtis and Reid [1972], Lasdon and Beck [1981], and Tomlin [1973]. Gill et al. [1981, 1984d, 1985] provide a good discussion on guidelines for model building and their influence on algorithms.

Finally, we mention that various modeling languages, such as GAMS (see Brooke et al., 1985), LINGO (see Cunningham and Schrage, 1989), and AMPL (see Fourer et al., 1990), are available to assist in the implementation of models and algorithms. Various nonlinear programming software packages, such as MINOS (see Murtagh and Saunders, 1982), GINO (see Liebman et al., 1986), GRG2 (see Lasdon et al., 1978), CONOPT (see Drud, 1985), SQP (see Mahidhara and Lasdon, 1990), LSGRG (see Smith and Lasdon, 1992), BARON (see Sahinidis, 1996), and LGO (see Pintér, 2000, 2001), among others, are also available to facilitate implementation. (The latter two are global optimizer software packages—see Chapter 11.) For a general discussion on algorithms and software evaluation for nonlinear optimization, see DiPillo and Murli [2003].

Part 1

Convex Analysis

This Page Intentionally Left Blank

Chapter 2 Convex Sets

The concept of convexity is of great importance in the study of optimization problems. Convex sets, polyhedral sets, and separation of disjoint convex sets are used frequently in the analysis of mathematical programming problems, the characterization of their optimal solutions, and in the development of computational procedures.

Following is an outline of the chapter. The reader is encouraged to review the mathematical preliminaries given in Appendix A.

Section 2.1: Convex Hulls This section is elementary. It presents some examples of convex sets and defines convex hulls. Readers having previous knowledge of convex sets may skip this section (with the possible exception of the Carathéodory theorem).

Section 2.2: Closure and Interior of a Set Some topological properties of sets related to interior, boundary, and closure points are discussed.

Section 2.3: Weierstrass's Theorem We discuss the concepts of min, max, inf, and sup and present an important result relating to the existence of minimizing or maximizing solutions.

Section 2.4: Separation and Support of Sets This section is important, since the notions of separation and support of convex sets are used frequently in optimization. A careful study of this section is recommended.

Section 2.5: Convex Cones and Polarity This short section dealing mainly with polar cones may be skipped without loss of continuity.

Section 2.6: Polyhedral Sets, Extreme Points, and Extreme Directions This section treats the special important case of polyhedral sets. Characterization of extreme points and extreme directions of polyhedral sets is developed. Also, the representation of a polyhedral set in terms of its extreme points and extreme directions is proved.

Section 2.7: Linear Programming and the Simplex Method The well-known simplex method is developed as a natural extension of the material in the preceding section. Readers who are familiar with the simplex method may skip this section. A polynomial-time algorithm for linear programming problems is discussed in Chapter 9.

2.1 Convex Hulls

In this section we first introduce the notions of convex sets and convex hulls. We then demonstrate that any point in the convex hull of a set S can be represented in terms of $n + 1$ points in the set S.

2.1.1 Definition

A set S in R^n is said to be *convex* if the *line segment* joining any two points of the set also belongs to the set. In other words, if $\mathbf{x}_1$ and $\mathbf{x}_2$ are in S, then $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$, must also belong to S for each $\lambda \in [0,1]$. Weighted averages of the form $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$, where $\lambda \in [0,1]$, are referred to as *convex combinations* of $\mathbf{x}_1$ and $\mathbf{x}_2$. Inductively, weighted averages of the form $\sum_{j=1}^{k} \lambda_j \mathbf{x}_j$, where $\sum_{j=1}^{k} \lambda_j = 1$, $\lambda_j \geq 0$, $j = 1,..., k$, are also called *convex combinations* of $\mathbf{x}_1,..., \mathbf{x}_k$. In this definition, if the nonnegativity conditions on the multipliers λ_j is dropped, $j = 1,..., k$, the combination is known as an *affine combination.* Finally, a combination $\sum_{j=1}^{k} \lambda_j \mathbf{x}_j$ where the multipliers λ_j, $j = 1,..., k$, are simply required to be in R, is known as a *linear combination.*

Figure 2.1 illustrates the notion of a convex set. Note that in Figure 2.1*b*, the line segment joining $\mathbf{x}_1$ and $\mathbf{x}_2$ does not lie entirely in the set.

The following are examples of convex sets:

1. $S = \{(x_1, x_2, x_3) : x_1 + 2x_2 - x_3 = 4\} \subset R^3$.

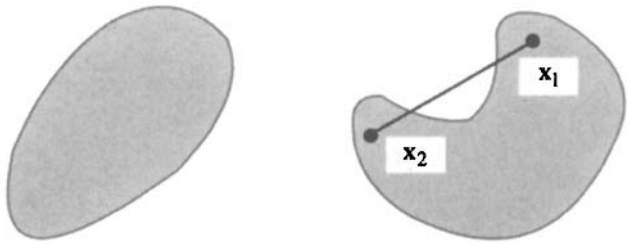
   This is an equation of a plane in R^3. In general, $S = \{\mathbf{x} : \mathbf{p}^t\mathbf{x} = \alpha\}$ is called a *hyperplane in* R^n, where $\mathbf{p}$ is a nonzero vector in R^n, usually referred to as the *gradient*, or *normal*, to the hyperplane, and α is a scalar. Note that if $\bar{\mathbf{x}} \in S$, we have $\mathbf{p}^t\bar{\mathbf{x}} = \alpha$, so that we can equivalently write $S = \{\mathbf{x} : \mathbf{p}^t(\mathbf{x} - \bar{\mathbf{x}}) = 0\}$. Hence, the vector $\mathbf{p}$ is orthogonal to all vectors $(\mathbf{x} - \bar{\mathbf{x}})$ for $\mathbf{x} \in S$, so it is perpendicular to the surface of the hyperplane S.

2. $S = \{(x_1, x_2, x_3) : x_1 + 2x_2 - x_3 \leq 4\} \subset R^3$.

   These are points on one side of the hyperplane defined above. These points form a *half-space.* In general, a half-space $S = \{\mathbf{x} : \mathbf{p}^t\mathbf{x} \leq \alpha\}$ in R^n is a convex set.

3. $S = \{(x_1, x_2, x_3) : x_1 + 2x_2 - x_3 \leq 4, \quad 2x_1 - x_2 + x_3 \leq 6\} \subset R^3$.

   This set is the intersection of two half-spaces. In general, the set $S = \{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ is a convex set, where $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b}$ is an m-vector. This set is the intersection of m half-spaces and is usually called a *polyhedral set.*



(*a*) Convex set (*b*) Nonconvex set

Figure 2.1 Convex and nonconvex sets.

4. $S = \{(x_1, x_2) : x_2 \geq |x_1|\} \subset R^2$.

   This set represents a *convex cone* in R^2 and is treated more fully in Section 2.4.

5. $S = \{(x_1, x_2) : x_1^2 + x_2^2 \leq 4\} \subset R^2$.

   This set represents points on and inside a circle with center (0, 0) and radius 2.

6. $S = \{\mathbf{x} : \mathbf{x}$ solves Problem P below$\}$:

$$\begin{aligned} \text{Problem P: Minimize } & \mathbf{c}^t\mathbf{x} \\ \text{subject to } & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned}$$

Here, $\mathbf{c}$ is an n-vector, $\mathbf{b}$ is an m-vector, $\mathbf{A}$ is an $m \times n$ matrix, and $\mathbf{x}$ is an n-vector. The set S gives all optimal solutions to the *linear programming problem* of minimizing the linear function $\mathbf{c}^t\mathbf{x}$ over the polyhedral region defined by $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$. This set itself happens to be a polyhedral set, being the intersection of $\mathbf{c}^t\mathbf{x} = v^*$ with $\mathbf{Ax} = \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$, where v^* is the optimal value of P.

The following lemma is an immediate consequence of the definition of convexity. It states that the intersection of two convex sets is convex and that the algebraic sum of two convex sets is also convex. The proof is elementary and is left as an exercise.

2.1.2 Lemma

Let S_1 and S_2 be convex sets in R^n. Then:

1. $S_1 \cap S_2$ is convex.
2. $S_1 \oplus S_2 = \{\mathbf{x}_1 + \mathbf{x}_2 : \mathbf{x}_1 \in S_1, \mathbf{x}_2 \in S_2\}$ is convex.
3. $S_1 \ominus S_2 = \{\mathbf{x}_1 - \mathbf{x}_2 : \mathbf{x}_1 \in S_1, \mathbf{x}_2 \in S_2\}$ is convex.

Convex Hulls

Given an arbitrary set S in R^n, different convex sets can be generated from S. In particular, we discuss below the convex hull of S.

2.1.3 Definition

Let S be an arbitrary set in R^n. The *convex hull* of S, denoted conv(S), is the collection of all convex combinations of S. In other words, $\mathbf{x} \in$ conv(S) if and only if $\mathbf{x}$ can be represented as

$$\mathbf{x} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j$$

$$\sum_{j=1}^{k} \lambda_j = 1$$

$$\lambda_j \geq 0 \qquad \text{for } j = 1,...,k,$$

where k is a positive integer and $\mathbf{x}_1,...,\mathbf{x}_k \in S$.

Figure 2.2 shows some examples of convex hulls. Actually, we see that in each case, conv(S) is the minimal (tightest enveloping) convex set that contains S. This is indeed the case in general, as given in Lemma 2.1.4. The proof is left as an exercise.

2.1.4 Lemma

Let S be an arbitrary set in R^n. Then, conv(S) is the smallest convex set containing S. Indeed, conv(S) is the intersection of all convex sets containing S.

Similar to the foregoing discussion, we can define the *affine hull* of S as the collection of all affine combinations of points in S. This is the smallest dimensional affine subspace that contains S. For example, the affine hull of two distinct points is the one-dimensional line containing these two points. Similarly, the *linear hull* of S is the collection of all linear combinations of points in S.

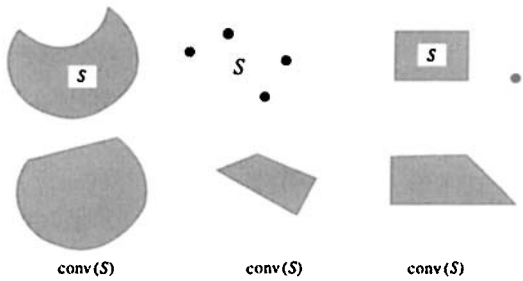


Figure 2.2 Convex hulls.

We have discussed above the convex hull of an arbitrary set S. The convex hull of a finite number of points leads to the definitions of a polytope and a simplex.

2.1.5 Definition

The convex hull of a finite number of points $\mathbf{x}_1,...,\mathbf{x}_{k+1}$ in R^n is called a *polytope*. If $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_k$, and $\mathbf{x}_{k+1}$ are *affinely independent*, which means that $\mathbf{x}_2-\mathbf{x}_1$, $\mathbf{x}_3-\mathbf{x}_1,...,\mathbf{x}_{k+1}-\mathbf{x}_1$ are linearly independent, then $\text{conv}(\mathbf{x}_1,...,\mathbf{x}_{k+1})$, the convex hull of $\mathbf{x}_1,...,\mathbf{x}_{k+1}$, is called a *simplex* having vertices $\mathbf{x}_1,...,\mathbf{x}_{k+1}$.

Figure 2.3 shows examples of a polytope and a simplex in R^n. Note that the maximum number of linearly independent vectors in R^n is n, and hence, there could be no simplex in R^n having more than $n + 1$ vertices.

Carathéodory Theorem

By definition, a point in the convex hull of a set can be represented as a convex combination of a finite number of points in the set. The following theorem shows that any point $\mathbf{x}$ in the convex hull of a set S can be represented as a convex combination of, at most, $n + 1$ points in S. The theorem is trivially true for $\mathbf{x} \in S$.

2.1.6 Theorem

Let S be an arbitrary set in R^n. If $\mathbf{x} \in \text{conv}(S)$, $\mathbf{x} \in \text{conv}(\mathbf{x}_1,...,\mathbf{x}_{n+1})$, where $\mathbf{x}_j \in S$ for $j = 1,..., n + 1$. In other words, $\mathbf{x}$ can be represented as

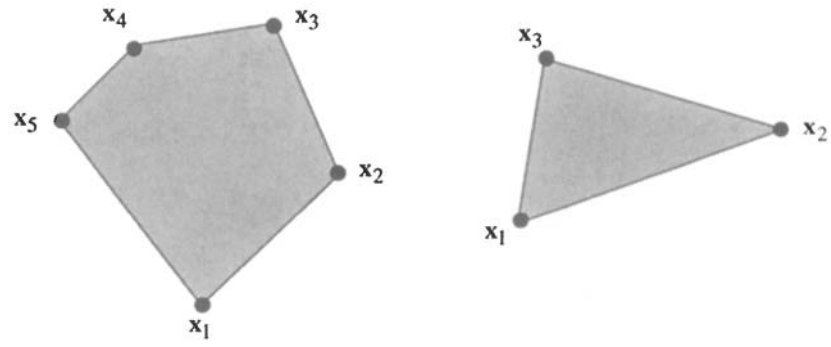


Figure 2.3 Polytope and simplex.

$$\mathbf{x} = \sum_{j=1}^{n+1} \lambda_j \mathbf{x}_j$$

$$\sum_{j=1}^{n+1} \lambda_j = 1$$

$$\lambda_j \geq 0 \qquad \text{for } j = 1,\ldots,n+1$$

$$\mathbf{x}_j \in S \qquad \text{for } j = 1,\ldots,n+1.$$

Proof

Since $\mathbf{x} \in \text{conv}(S)$, $\mathbf{x} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j$, where $\lambda_j > 0$ for $j = 1,\ldots, k$, $\mathbf{x}_j \in S$ for $j = 1,\ldots, k$, and $\sum_{j=1}^{k} \lambda_j = 1$. If $k \leq n+1$, the result is at hand. Now suppose that $k > n+1$. A reader familiar with basic feasible solutions and extreme points (see Theorem 2.5.4) will now notice immediately that at an extreme point of the set $\{\lambda : \sum_{j=1}^{k} \lambda_j \mathbf{x}_j = \mathbf{x}, \sum_{j=1}^{k} \lambda_j = 1, \lambda \geq \mathbf{0}\}$, no more than $n + 1$ components of λ are positive, hence proving the result. However, let us continue to provide an independent argument.

Toward this end, note that $\mathbf{x}_2 - \mathbf{x}_1$, $\mathbf{x}_3 - \mathbf{x}_1,\ldots,\mathbf{x}_k - \mathbf{x}_1$ are linearly dependent. Thus, there exist scalars μ_2, $\mu_3,\ldots,\mu_k$ not all zero such that $\sum_{j=2}^{k} \mu_j(\mathbf{x}_j - \mathbf{x}_1) = \mathbf{0}$. Letting $\mu_1 = -\sum_{j=2}^{k} \mu_j$, it follows that $\sum_{j=1}^{k} \mu_j \mathbf{x}_j = \mathbf{0}$, $\sum_{j=1}^{k} \mu_j = 0$, and not all the μ_j values are equal to zero. Note that at least one μ_j is larger than zero. Then

$$\mathbf{x} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j + \mathbf{0} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j - \alpha \sum_{j=1}^{k} \mu_j \mathbf{x}_j = \sum_{j=1}^{k} (\lambda_j - \alpha\mu_j)\mathbf{x}_j$$

for any real α. Now, choose α as follows:

$$\alpha = \underset{1 \leq j \leq k}{\text{minimum}} \left\{ \frac{\lambda_j}{\mu_j} : \mu_j > 0 \right\} = \frac{\lambda_i}{\mu_i} \qquad \text{for some } i \in \{1,\ldots,k\}.$$

Note that $\alpha > 0$. If $\mu_j \leq 0$, $\lambda_j - \alpha\mu_j > 0$, and if $\mu_j > 0$, $\lambda_j/\mu_j \geq \lambda_i/\mu_i = \alpha$ and hence $\lambda_j - \alpha\mu_j \geq 0$. In other words, $\lambda_j - \alpha\mu_j \geq 0$ for all $j = 1,\ldots, k$. In particular, $\lambda_i - \alpha\mu_i = 0$ by the definition of α. Therefore, $\mathbf{x} = \sum_{j=1}^{k} (\lambda_j - \alpha\mu_j)\mathbf{x}_j$, where $\lambda_j - \alpha\mu_j \geq 0$ for $j = 1,\ldots, k$, $\sum_{j=1}^{k} (\lambda_j - \alpha\mu_j) = 1$, and furthermore, $\lambda_i - \alpha\mu_i = 0$. Consequently, we have represented $\mathbf{x}$ as a convex combination of at most $k - 1$ points in S. This process can be repeated until $\mathbf{x}$ is represented as a convex combination of at most $n + 1$ points in S. This completes the proof.

2.2 Closure and Interior of a Set

In this section we develop some topological properties of sets in general and of convex sets in particular. As a preliminary, given a point $\mathbf{x}$ in R^n, an ε-neighborhood around it is the set $N_\varepsilon(\mathbf{x}) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| < \varepsilon\}$. Let us first review the definitions of closure, interior, and boundary of an arbitrary set in R^n, using the concept of an ε-neighborhood.

2.2.1 Definition

Let S be an arbitrary set in R^n. A point $\mathbf{x}$ is said to be in the *closure* of S, denoted by cl S, if $S \cap N_\varepsilon(\mathbf{x}) \neq \varnothing$ for every $\varepsilon > 0$. If $S = \text{cl } S$, S is called *closed.* A point $\mathbf{x}$ is said to be in the *interior* of S, denoted int S, if $N_\varepsilon(\mathbf{x}) \subset S$ for some $\varepsilon > 0$. A *solid set* $S \subseteq R^n$ is one having a nonempty interior. If $S = \text{int } S$, S is called *open.* Finally, $\mathbf{x}$ is said to be in the *boundary* of S, denoted ∂S, if $N_\varepsilon(\mathbf{x})$ contains at least one point in S and one point not in S for every $\varepsilon > 0$. A set S is *bounded* if it can be contained in a ball of a sufficiently large radius. A *compact set* is one that is both closed and bounded. Note that the complement of an open set is a closed set (and vice versa), and that the boundary points of any set and its complement are the same.

To illustrate, consider $S = \{(x_1, x_2) : x_1^2 + x_2^2 \leq 1\}$, which represents all points within a circle with center (0, 0) and radius 1. It can easily be verified that S is closed; that is, S = cl S. Furthermore, int S consists of all points that lie strictly within the circle; that is, int $S = \{(x_1, x_2) : x_1^2 + x_2^2 < 1\}$. Finally, ∂S consists of points on the circle; that is, $\partial S = \{(x_1, x_2) : x_1^2 + x_2^2 = 1\}$.

Hence, a set S is closed if and only if it contains all its boundary points (i.e., $\partial S \subseteq S$). Moreover, cl $S \equiv S \cup \partial S$ is the smallest closed set containing S. Similarly, a set is open if and only if it does not contain any of its boundary points (more precisely, $\partial S \cap S = \varnothing$). Clearly, a set may be neither open nor closed, and the only sets in R^n that are both open and closed are the empty set and R^n itself. Also, note that any point $\mathbf{x} \in S$ must be either an interior or a boundary point of S. However, $S \neq \text{int } S \cup \partial S$, since S need not contain its boundary points. But since int $S \subseteq S$, we have int $S = S - \partial S$, while $\partial S \neq S - \text{int } S$ necessarily.

There is another equivalent definition of a closed set, which is often important from the viewpoint of demonstrating that a set is closed. This definition is based on sequences of points contained in S (review Appendix A for related mathematical concepts). A set S is closed if and only if for any convergent sequence of points $\{\mathbf{x}_k\}$ contained in S with limit point $\bar{\mathbf{x}}$, we also have that $\bar{\mathbf{x}} \in S$. The equivalence of this and the previous definition of

closedness is easily seen by noting that the limit point $\bar{\mathbf{x}}$ of any convergent sequence of points in S must either lie in the interior or on the boundary of S, since otherwise, there would exist an $\varepsilon > 0$ such that $\{\mathbf{x} : \|\mathbf{x} - \bar{\mathbf{x}}\| < \varepsilon\} \cap S = \varnothing$, contradicting that $\bar{\mathbf{x}}$ is the limit point of a sequence contained in S. Hence, if S is closed, $\bar{\mathbf{x}} \in S$. Conversely, if S satisfies the sequence property above, it is closed, since otherwise there would exist some boundary point $\bar{\mathbf{x}}$ not contained in S. But by the definition of a boundary point, the set $N_{\varepsilon^k}(\bar{\mathbf{x}}) \cap S \neq \varnothing$ for each $k = 1, 2,...,$ where $0 < \varepsilon < 1$ is some scalar. Hence, selecting some $\mathbf{x}_k \in N_{\varepsilon^k}(\bar{\mathbf{x}}) \cap S$ for each $k = 1, 2,...,$ we will have $\{\mathbf{x}_k\} \subseteq S$; and clearly $\{\mathbf{x}_k\} \to \bar{\mathbf{x}}$, which means that we must have $\bar{\mathbf{x}} \in S$ by our hypothesis. This is a contradiction.

To illustrate, note that the polyhedral set $S = \{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ is closed, since given any convergent sequence $\{\mathbf{x}_k\} \subseteq S$, with $\{\mathbf{x}_k\} \to \bar{\mathbf{x}}$, we also have $\bar{\mathbf{x}} \in S$. This follows because $\mathbf{A}\mathbf{x}_k \leq \mathbf{b}$ for all k; so by the continuity of linear functions, we have in the limit that $\mathbf{A}\bar{\mathbf{x}} \leq \mathbf{b}$ as well, or that $\bar{\mathbf{x}} \in S$.

Line Segment Between Points in the Closure and the Interior of a Set

Given a convex set having a nonempty interior, the line segment (excluding the endpoints) joining a point in the interior of the set and a point in the closure of the set belongs to the interior of the set. This result is proved below. (Exercise 2.43 suggests a means for constructing a simpler proof based on the concept of supporting hyperplanes introduced in Section 2.4.)

2.2.2 Theorem

Let S be a convex set in R^n with a nonempty interior. Let $\mathbf{x}_1 \in \text{cl } S$ and $\mathbf{x}_2 \in \text{int } S$. Then $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in \text{int } S$ for each $\lambda \in (0,1)$.

Proof

Since $\mathbf{x}_2 \in \text{int } S$, there exists an $\varepsilon > 0$ such that $\{\mathbf{z} : \|\mathbf{z} - \mathbf{x}_2\| < \varepsilon\} \subset S$. Let $\mathbf{y}$ be such that

$$\mathbf{y} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2, \tag{2.1}$$

where $\lambda \in (0,1)$. To prove that $\mathbf{y}$ belongs to int S, it suffices to construct a neighborhood about $\mathbf{y}$ that also belongs to S. In particular, we show that $\{\mathbf{z} : \|\mathbf{z} - \mathbf{y}\| < (1-\lambda)\varepsilon\} \subset S$. Let $\mathbf{z}$ be such that $\|\mathbf{z} - \mathbf{y}\| < (1-\lambda)\varepsilon$ (refer to Figure 2.4). Since $\mathbf{x}_1 \in \text{cl } S$,

$$\left\{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_1\| < \frac{(1-\lambda)\varepsilon - \|\mathbf{z} - \mathbf{y}\|}{\lambda}\right\} \cap S$$

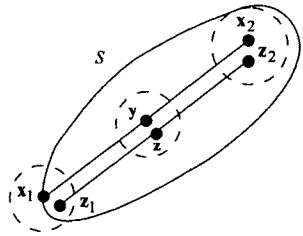


Figure 2.4 Line segment joining points in the closure and interior of a set.

is not empty. In particular, there exists a $\mathbf{z}_1 \in S$ such that

$$\|\mathbf{z}_1 - \mathbf{x}_1\| < \frac{(1-\lambda)\varepsilon - \|\mathbf{z} - \mathbf{y}\|}{\lambda}. \tag{2.2}$$

Now let $\mathbf{z}_2 = \dfrac{\mathbf{z} - \lambda\mathbf{z}_1}{1-\lambda}$. From (2.1), the Schwartz inequality, and (2.2), we get

$$\begin{aligned}\|\mathbf{z}_2 - \mathbf{x}_2\| = \left\|\frac{\mathbf{z} - \lambda\mathbf{z}_1}{1-\lambda} - \mathbf{x}_2\right\| &= \left\|\frac{(\mathbf{z} - \lambda\mathbf{z}_1) - (\mathbf{y} - \lambda\mathbf{x}_1)}{1-\lambda}\right\| \\ &= \frac{1}{1-\lambda}\|(\mathbf{z} - \mathbf{y}) + \lambda(\mathbf{x}_1 - \mathbf{z}_1)\| \\ &\le \frac{1}{1-\lambda}\left(\|\mathbf{z} - \mathbf{y}\| + \lambda\|\mathbf{x}_1 - \mathbf{z}_1\|\right) \\ &< \varepsilon.\end{aligned}$$

Therefore, $\mathbf{z}_2 \in S$. By the definition of $\mathbf{z}_2$, note that $\mathbf{z} = \lambda\mathbf{z}_1 + (1-\lambda)\mathbf{z}_2$; and since both $\mathbf{z}_1$ and $\mathbf{z}_2$ belong to S, $\mathbf{z}$ also belongs to S. We have shown that any $\mathbf{z}$ with $\|\mathbf{z} - \mathbf{y}\| < (1-\lambda)\varepsilon$ belongs to S. Therefore, $\mathbf{y} \in \operatorname{int} S$ and the proof is complete.

Corollary 1

Let S be a convex set. Then $\operatorname{int} S$ is convex.

Corollary 2

Let S be a convex set with a nonempty interior. Then $\operatorname{cl} S$ is convex.

Proof

Let $\mathbf{x}_1, \mathbf{x}_2 \in \operatorname{cl} S$. Pick $\mathbf{z} \in \operatorname{int} S$ (by assumption, $\operatorname{int} S \neq \varnothing$). By the theorem, $\lambda\mathbf{x}_2 + (1-\lambda)\mathbf{z} \in \operatorname{int} S$ for each $\lambda \in (0,1)$. Now fix $\mu \in (0,1)$. By the theorem, $\mu\mathbf{x}_1 + (1-\mu)[\lambda\mathbf{x}_2 + (1-\lambda)\mathbf{z}] \in \operatorname{int} S \subset S$ for each $\lambda \in (0,1)$. If we take the limit as λ approaches 1, it follows that $\mu\mathbf{x}_1 + (1-\mu)\mathbf{x}_2 \in \operatorname{cl} S$, and the proof is complete.

Corollary 3

Let S be a convex set with a nonempty interior. Then cl(int S) = cl S.

Proof

Clearly, cl(int S) $\subseteq$ cl S. Now let $\mathbf{x} \in \text{cl}\, S$, and pick $\mathbf{y} \in \text{int}\, S$ (by assumption, int $S \neq \varnothing$). Then $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in \text{int}\, S$ for each $\lambda \in (0,1)$. Letting $\lambda \to 1^-$, it follows that $\mathbf{x} \in$ cl(int S).

Corollary 4

Let S be a convex set with a nonempty interior. Then int(cl S) = int S.

Proof

Note that int $S \subseteq$ int (cl S). Let $\mathbf{x}_1 \in$ int(cl S). We need to show that $\mathbf{x}_1 \in$ int S. There exists an $\varepsilon > 0$ such that $\|\mathbf{y} - \mathbf{x}_1\| < \varepsilon$ implies that $\mathbf{y} \in$ cl S. Now let $\mathbf{x}_2 \neq \mathbf{x}_1$ belong to int S and let $\mathbf{y} = (1+\Delta)\mathbf{x}_1 - \Delta\mathbf{x}_2$, where $\Delta = \varepsilon / (2\|\mathbf{x}_1 - \mathbf{x}_2\|)$. Since $\|\mathbf{y} - \mathbf{x}_1\| = \varepsilon/2$, $\mathbf{y} \in$ cl S. But $\mathbf{x}_1 = \lambda\mathbf{y} + (1-\lambda)\mathbf{x}_2$, where $\lambda = 1/(1+\Delta) \in (0, 1)$. Since $\mathbf{y} \in$ cl S and $\mathbf{x}_2 \in$ int S, then, by the theorem, $\mathbf{x}_1 \in$ int S, and the proof is complete.

Theorem 2.2.2 and its corollaries can be strengthened considerably by using the notion of relative interiors (see the Notes and References section at the end of the chapter).

2.3 Weierstrass's Theorem

A very important and widely used result is based on the foregoing concepts. This result relates to the existence of a minimizing solution for an optimization problem. Here we say that $\bar{\mathbf{x}}$ is a minimizing solution for the problem $\min\{f(\mathbf{x}) : \mathbf{x} \in S\}$, provided that $\bar{\mathbf{x}} \in S$ and $f(\bar{\mathbf{x}}) \leq f(\mathbf{x})$ for all $\mathbf{x} \in S$. In such a case, we say that a minimum exists. On the other hand, we say that α = infimum$\{f(\mathbf{x}) : \mathbf{x} \in S\}$ (abbreviated *inf*) if α is the greatest lower bound of f on S; that is, $\alpha \leq f(\mathbf{x})$ for all $\mathbf{x} \in S$ and there is no $\bar{\alpha} > \alpha$ such that $\bar{\alpha} \leq f(\mathbf{x})$ for all $\mathbf{x} \in S$. Similarly, $\alpha = \max\{f(\mathbf{x}) : \mathbf{x} \in S\}$ if there exists a solution $\bar{\mathbf{x}} \in S$ such that $\alpha = f(\bar{\mathbf{x}}) \geq f(\mathbf{x})$ for all $\mathbf{x} \in S$. On the other hand, $\alpha = \text{supremum}\{f(\mathbf{x}) : \mathbf{x} \in S\}$ (abbreviated *sup*) if α is the least upper bound of f on S; that is, $\alpha \geq f(\mathbf{x})$ for all $\mathbf{x} \in S$, and there is no $\bar{\alpha} < \alpha$ such that $\bar{\alpha} \geq f(\mathbf{x})$ for all $\mathbf{x} \in S$.

Figure 2.5 illustrates three instances where a minimum does not exist. In Figure 2.5*a*, the infimum of f over (a, b) is given by $f(b)$, but since S is not closed and, in particular, $b \notin S$, a minimum does not exist. In Figure 2.5*b* we have that $\inf\{f(x) : x \in [a,b]\}$ is given by the limit of $f(x)$ as x approaches b

from the "left," denoted $\lim_{x \to b^-} f(x)$. However, since f is discontinuous at b, a minimizing solution does not exist. Finally, Figure 2.5*c* illustrates a situation in which f is unbounded over the unbounded set $S = \{x : x \geq a\}$.

We now formally state and prove the result that if S is nonempty, closed, and bounded, and if f is continuous on S, then unlike the various situations of Figure 2.5, a minimum exists. The reader is encouraged to study how these different assumptions guarantee the different assertions made in the following proof.

2.3.1 Theorem

Let S be a nonempty, compact set, and let f: $S \to R$ be continuous on S. Then the problem $\min\{f(\mathbf{x}) : \mathbf{x} \in S\}$ attains its minimum; that is, there exists a minimizing solution to this problem.

Proof

Since f is continuous on S and S is both closed and bounded, f is bounded below on S. Consequently, since $S \neq \varnothing$, there exists a greatest lower bound $\alpha \equiv \inf\{f(\mathbf{x}) : \mathbf{x} \in S\}$. Now let $0 < \varepsilon < 1$, and consider the set $S_k = \{\mathbf{x} \in S : \alpha \leq f(\mathbf{x}) \leq \alpha + \varepsilon^k\}$ for each $k = 1, 2, \ldots$. By the definition of an infimum, $S_k \neq \varnothing$ for each k, so we may construct a sequence of points $\{\mathbf{x}_k\} \subseteq S$ by selecting a point $\mathbf{x}_k \in S_k$ for each $k = 1, 2, \ldots$. Since S is bounded, there exists a convergent subsequence $\{\mathbf{x}_k\}_K \to \bar{\mathbf{x}}$, indexed by the set K. By the closedness of S, we have $\bar{\mathbf{x}} \in S$; and by the continuity of f, since $\alpha \leq f(\mathbf{x}_k) \leq \alpha + \varepsilon^k$ for all k, we have that $\alpha = \lim_{k \to \infty, k \in K} f(\mathbf{x}_k) = f(\bar{\mathbf{x}})$. Hence, we have shown that there exists a solution $\bar{\mathbf{x}} \in S$ such that $f(\bar{\mathbf{x}}) = \alpha = \inf\{f(\mathbf{x}) : \mathbf{x} \in S\}$, so $\bar{\mathbf{x}}$ is a minimizing solution. This completes the proof.

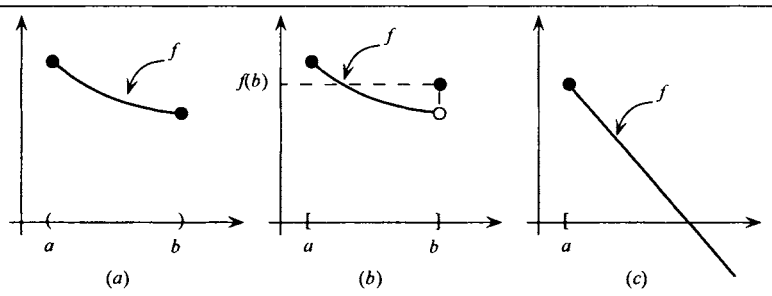


Figure 2.5 Nonexistence of a minimizing solution.

2.4 Separation and Support of Sets

The notions of supporting hyperplanes and separation of disjoint convex sets are very important in optimization. Almost all optimality conditions and duality relationships use some sort of separation or support of convex sets. The results of this section are based on the following geometric fact: Given a closed convex set S and a point $\mathbf{y} \notin S$, there exists a unique point $\bar{\mathbf{x}} \in S$ with minimum distance from $\mathbf{y}$ and a hyperplane that separates $\mathbf{y}$ and S.

Minimum Distance from a Point to a Convex Set

To establish the above important result, the following *parallelogram law* is needed. Let $\mathbf{a}$ and $\mathbf{b}$ be two vectors in R^n. Then

$$\|\mathbf{a}+\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\mathbf{a}^t\mathbf{b}$$

$$\|\mathbf{a}-\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\mathbf{a}^t\mathbf{b}.$$

By adding we get

$$\|\mathbf{a}+\mathbf{b}\|^2 + \|\mathbf{a}-\mathbf{b}\|^2 = 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2.$$

This result is illustrated in Figure 2.6 and can be interpreted as follows: The sum of squared norms of the diagonals of a parallelogram is equal to the sum of squared norms of its sides.

We now state and prove the *closest-point theorem*. Again, the reader is encouraged to investigate how the various assumptions play a role in guaranteeing the various assertions.

2.4.1 Theorem

Let S be a nonempty, closed convex set in R^n and $\mathbf{y} \notin S$. Then there exists a unique point $\bar{\mathbf{x}} \in S$ with minimum distance from $\mathbf{y}$. Furthermore, $\bar{\mathbf{x}}$ is the minimizing point if and only if $(\mathbf{y}-\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}}) \le 0$ for all $\mathbf{x} \in S$.

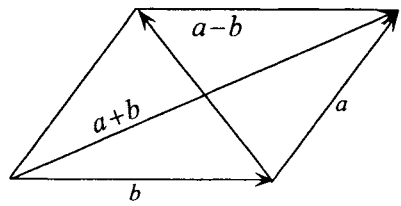


Figure 2.6 Parallelogram law.

Proof

First, let us establish the existence of a closest point. Since $S \neq \emptyset$, there exists a point $\hat{\mathbf{x}} \in S$, and we can confine our attention to the set $\bar{S} = S \cap \{\mathbf{x}: \|\mathbf{y}-\mathbf{x}\| \leq \|\mathbf{y}-\hat{\mathbf{x}}\|\}$ in seeking the closest point. In other words, the closest-point problem $\inf\{\|\mathbf{y}-\mathbf{x}\|: \mathbf{x} \in S\}$ is equivalent to $\inf\{\|\mathbf{y}-\mathbf{x}\|: \mathbf{x} \in \bar{S}\}$. But the latter problem involves finding the minimum of a continuous function over a nonempty, compact set $\bar{S}$, so by Weierstrass's theorem, Theorem 2.3.1, we know that there exists a minimizing point $\bar{\mathbf{x}}$ in S that is closest to the point $\mathbf{y}$.

To show uniqueness, suppose that there is an $\bar{\mathbf{x}}' \in S$ such that $\|\mathbf{y}-\bar{\mathbf{x}}\| = \|\mathbf{y}-\bar{\mathbf{x}}'\| = \gamma$. By the convexity of S, $(\bar{\mathbf{x}}+\bar{\mathbf{x}}')/2 \in S$. By the triangle inequality we get

$$\left\|\mathbf{y}-\frac{\bar{\mathbf{x}}+\bar{\mathbf{x}}'}{2}\right\| \leq \frac{1}{2}\|\mathbf{y}-\bar{\mathbf{x}}\| + \frac{1}{2}\|\mathbf{y}-\bar{\mathbf{x}}'\| = \gamma.$$

If strict inequality holds, we have a contradiction to $\bar{\mathbf{x}}$ being the closest point to $\mathbf{y}$. Therefore, equality holds, and we must have $\mathbf{y}-\bar{\mathbf{x}} = \lambda(\mathbf{y}-\bar{\mathbf{x}}')$ for some λ. Since $\|\mathbf{y}-\bar{\mathbf{x}}\| = \|\mathbf{y}-\bar{\mathbf{x}}'\| = \gamma$, we have $|\lambda| = 1$. Clearly, $\lambda \neq -1$, because otherwise, $\mathbf{y} = (\bar{\mathbf{x}}+\bar{\mathbf{x}}')/2 \in S$, contradicting the assumption that $\mathbf{y} \notin S$. So $\lambda = 1$, yielding $\bar{\mathbf{x}}' = \bar{\mathbf{x}}$, and uniqueness is established.

To complete the proof, we need to show that $(\mathbf{y}-\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}}) \leq 0$ for all $\mathbf{x} \in S$ is both a necessary and a sufficient condition for $\bar{\mathbf{x}}$ to be the point in S closest to $\mathbf{y}$.

To prove sufficiency, let $\mathbf{x} \in S$. Then

$$\|\mathbf{y}-\mathbf{x}\|^2 = \|\mathbf{y}-\bar{\mathbf{x}}+\bar{\mathbf{x}}-\mathbf{x}\|^2 = \|\mathbf{y}-\bar{\mathbf{x}}\|^2 + \|\bar{\mathbf{x}}-\mathbf{x}\|^2 + 2(\bar{\mathbf{x}}-\mathbf{x})^t(\mathbf{y}-\bar{\mathbf{x}}).$$

Since $\|\bar{\mathbf{x}}-\mathbf{x}\|^2 \geq 0$ and $(\bar{\mathbf{x}}-\mathbf{x})^t(\mathbf{y}-\bar{\mathbf{x}}) \geq 0$ by assumption, $\|\mathbf{y}-\mathbf{x}\|^2 \geq \|\mathbf{y}-\bar{\mathbf{x}}\|^2$ and $\bar{\mathbf{x}}$ is the minimizing point. Conversely, assume that $\|\mathbf{y}-\mathbf{x}\|^2 \geq \|\mathbf{y}-\bar{\mathbf{x}}\|^2$ for all $\mathbf{x} \in S$. Let $\mathbf{x} \in S$ and note that $\bar{\mathbf{x}} + \lambda(\mathbf{x}-\bar{\mathbf{x}}) \in S$ for $0 \leq \lambda \leq 1$ by the convexity of S. Therefore,

$$\|\mathbf{y}-\bar{\mathbf{x}}-\lambda(\mathbf{x}-\bar{\mathbf{x}})\|^2 \geq \|\mathbf{y}-\bar{\mathbf{x}}\|^2. \tag{2.3}$$

Also,

$$\|\mathbf{y}-\bar{\mathbf{x}}-\lambda(\mathbf{x}-\bar{\mathbf{x}})\|^2 = \|\mathbf{y}-\bar{\mathbf{x}}\|^2 + \lambda^2(\mathbf{x}-\bar{\mathbf{x}}^2) - 2\lambda(\mathbf{y}-\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}}). \tag{2.4}$$

From (2.3) and (2.4) we get

$$2\lambda(\mathbf{y}-\bar{\mathbf{x}})^t(\mathbf{x}-\bar{\mathbf{x}}) \leq \lambda^2\|\mathbf{x}-\bar{\mathbf{x}}\|^2 \tag{2.5}$$

for all $0 \le \lambda \le 1$. Dividing (2.5) by any such $\lambda > 0$ and letting $\lambda \to 0^+$, the result follows.

Theorem 2.4.1 is illustrated in Figure 2.7*a*. Note that the angle between $(\mathbf{y} - \bar{\mathbf{x}})$ and $(\mathbf{x} - \bar{\mathbf{x}})$ for any point $\mathbf{x}$ in S is greater than or equal to 90°, and hence $(\mathbf{y} - \bar{\mathbf{x}})^t(\mathbf{x} - \bar{\mathbf{x}}) \le 0$ for all $\mathbf{x} \in S$. This says that the set S lies in the half-space $\alpha^t(\mathbf{x} - \bar{\mathbf{x}}) \le 0$ relative to the hyperplane $\alpha^t(\mathbf{x} - \bar{\mathbf{x}}) = 0$ passing through $\bar{\mathbf{x}}$ and having a normal $\alpha = (\mathbf{y} - \bar{\mathbf{x}})$. Note also by referring to Figure 2.7*b* that this feature does not necessarily hold even over $N_\varepsilon(\bar{\mathbf{x}}) \cap S$ if S is not convex.

Hyperplanes and Separation of Two Sets

Since we shall be dealing with separating and supporting hyperplanes, precise definitions of hyperplanes and half-spaces are reiterated below.

2.4.2 Definition

A *hyperplane* H in R^n is a collection of points of the form $\{\mathbf{x} : \mathbf{p}^t\mathbf{x} = \alpha\}$, where $\mathbf{p}$ is a nonzero vector in R^n and α is a scalar. The vector $\mathbf{p}$ is called the *normal* vector of the hyperplane. A hyperplane H defines two *closed half-spaces* $H^+ = \{\mathbf{x} : \mathbf{p}^t\mathbf{x} \ge \alpha\}$ and $H^- = \{\mathbf{x} : \mathbf{p}^t\mathbf{x} \le \alpha\}$ and the two *open half-spaces* $\{\mathbf{x} : \mathbf{p}^t\mathbf{x} > \alpha\}$ and $\{\mathbf{x} : \mathbf{p}^t\mathbf{x} < \alpha\}$.

Note that any point in R^n lies in H^+, in H^-, or in both. Also, a hyperplane H and the corresponding half-spaces can be written in reference to a fixed point, say, $\bar{\mathbf{x}} \in H$. If $\bar{\mathbf{x}} \in H$, $\mathbf{p}^t\bar{\mathbf{x}} = \alpha$ and hence any point $\mathbf{x} \in H$ must satisfy $\mathbf{p}^t\mathbf{x} - \mathbf{p}^t\bar{\mathbf{x}} = \alpha - \alpha = 0$; that is, $\mathbf{p}^t(\mathbf{x} - \bar{\mathbf{x}}) = 0$. Accordingly, $H^+ = \{\mathbf{x} : \mathbf{p}^t(\mathbf{x} - \bar{\mathbf{x}}) \ge 0\}$ and $H^- = \{\mathbf{x} : \mathbf{p}^t(\mathbf{x} - \bar{\mathbf{x}}) \le 0\}$. Figure 2.8 shows a hyperplane H passing through $\bar{\mathbf{x}}$ and having a normal vector $\mathbf{p}$.

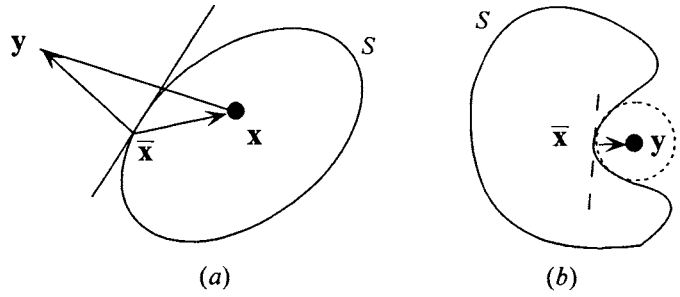


Figure 2.7 Minimum distance to a closed convex set.

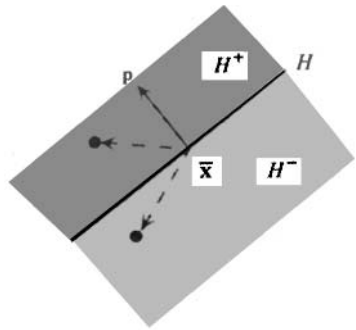


Figure 2.8 Hyperplane and corresponding half-spaces.

As an example, consider $H = \{(x_1, x_2, x_3, x_4) : x_1 + x_2 - x_3 + 2x_4 = 4\}$. The normal vector is $\mathbf{p} = (1,1,-1,2)^t$. Alternatively, the hyperplane can be written in reference to any point in H: for example, $\bar{\mathbf{x}} = (0,6,0,-1)^t$. In this case we write $H = \{(x_1, x_2, x_3, x_4) : x_1 + (x_2 - 6) - x_3 + 2(x_4 + 1) = 0\}$.

2.4.3 Definition

Let S_1 and S_2 be nonempty sets in R^n. A hyperplane $H = \{\mathbf{x} : \mathbf{p}^t\mathbf{x} = \alpha\}$ is said to *separate* S_1 and S_2 if $\mathbf{p}^t\mathbf{x} \geq \alpha$ for each $\mathbf{x} \in S_1$ and $\mathbf{p}^t\mathbf{x} \leq \alpha$ for each $\mathbf{x} \in S_2$. If, in addition, $S_1 \cup S_2 \not\subset H$, H is said to *properly separate* S_1 and S_2. The hyperplane H is said to *strictly separate* S_1 and S_2 if $\mathbf{p}^t\mathbf{x} > \alpha$ for each $\mathbf{x} \in S_1$ and $\mathbf{p}^t\mathbf{x} < \alpha$ for each $\mathbf{x} \in S_2$. The hyperplane H is said to *strongly separate* S_1 and S_2 if $\mathbf{p}^t\mathbf{x} \geq \alpha + \varepsilon$ for each $\mathbf{x} \in S_1$ and $\mathbf{p}^t\mathbf{x} \leq \alpha$ for each $\mathbf{x} \in S_2$, where ε is a positive scalar.

Figure 2.9 shows various types of separation. Of course, strong separation implies strict separation, which implies proper separation, which in turn implies separation. *Improper separation* is usually of little value, since it corresponds to a hyperplane containing both S_1 and S_2, as shown in Figure 2.9.

Separation of a Convex Set and a Point

We shall now present the first and most fundamental separation theorem. Other separation and support theorems will follow from this basic result.

2.4.4 Theorem

Let S be a nonempty closed convex set in R^n and $\mathbf{y} \notin S$. Then there exists a nonzero vector $\mathbf{p}$ and a scalar α such that $\mathbf{p}^t\mathbf{y} > \alpha$ and $\mathbf{p}^t\mathbf{x} \leq \alpha$ for each $\mathbf{x} \in S$.

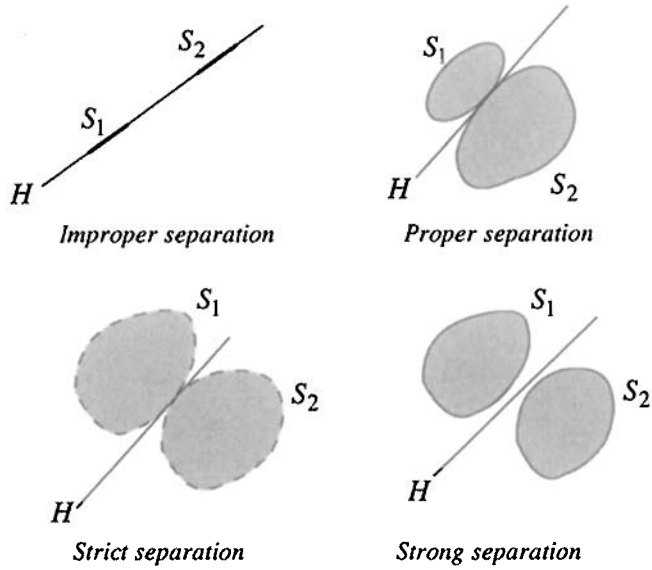


Figure 2.9 Various types of separation.

Proof

The set S is a nonempty closed convex set and $\mathbf{y} \notin S$. Hence, by Theorem 2.4.1, there exists a unique minimizing point $\bar{\mathbf{x}} \in S$ such that $(\mathbf{x} - \bar{\mathbf{x}})^t (\mathbf{y} - \bar{\mathbf{x}}) \leq 0$ for each $\mathbf{x} \in S$.

Letting $\mathbf{p} = \mathbf{y} - \bar{\mathbf{x}} \neq 0$ and $\alpha = \bar{\mathbf{x}}^t (\mathbf{y} - \bar{\mathbf{x}}) = \mathbf{p}^t \bar{\mathbf{x}}$, we get $\mathbf{p}^t \mathbf{x} \leq \alpha$ for each $\mathbf{x} \in S$, while $\mathbf{p}^t \mathbf{y} - \alpha = (\mathbf{y} - \bar{\mathbf{x}})^t (\mathbf{y} - \bar{\mathbf{x}}) = \|\mathbf{y} - \bar{\mathbf{x}}\|^2 > 0$. This completes the proof.

Corollary 1

Let S be a closed convex set in R^n. Then S is the intersection of all half-spaces containing S.

Proof

Obviously, S is contained in the intersection of all half-spaces containing it. In contradiction of the desired result, suppose that there is a point $\mathbf{y}$ in the intersection of these half-spaces but not in S. By the theorem, there exists a half-space that contains S but not $\mathbf{y}$. This contradiction proves the corollary.

Corollary 2

Let S be a nonempty set, and let $\mathbf{y} \notin \text{cl conv}(S)$, the closure of the convex hull of S. Then there exists a strongly separating hyperplane for S and $\mathbf{y}$.

Proof

The result follows by letting cl conv(S) play the role of S in Theorem 2.4.4.

The following statements are equivalent to the conclusion of the theorem. The reader is asked to verify this equivalence. Note that statements 1 and 2 are equivalent only in this special case since $\mathbf{y}$ is a point. Note also that in Theorem 2.4.4, we have $\alpha \equiv \mathbf{p}^t\bar{\mathbf{x}} = \max\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S\}$, since for any $\mathbf{x} \in S$, $\mathbf{p}^t(\bar{\mathbf{x}} - \mathbf{x}) = (\mathbf{y} - \bar{\mathbf{x}})^t(\bar{\mathbf{x}} - \mathbf{x}) \geq 0$.

1. There exists a hyperplane that *strictly* separates S and $\mathbf{y}$.
2. There exists a hyperplane that *strongly* separates S and $\mathbf{y}$.
3. There exists a vector $\mathbf{p}$ such that $\mathbf{p}^t\mathbf{y} > \sup\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S\}$.
4. There exists a vector $\mathbf{p}$ such that $\mathbf{p}^t\mathbf{y} < \inf\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S\}$.

Farkas's Theorem as a Consequence of Theorem 2.4.4

Farkas's Theorem is used extensively in the derivation of optimality conditions of linear and nonlinear programming problems. The theorem can be stated as follows. Let $\mathbf{A}$ be an $m \times n$ matrix, and let $\mathbf{c}$ be an n-vector. Then exactly one of the following two systems has a solution:

***System* 1**: $\mathbf{Ax} \leq \mathbf{0}$ and $\mathbf{c}^t\mathbf{x} > 0$ for some $\mathbf{x} \in R^n$.

***System* 2**: $\mathbf{A}^t\mathbf{y} = \mathbf{c}$ and $\mathbf{y} \geq \mathbf{0}$ for some $\mathbf{y} \in R^m$.

If we denote the columns of $\mathbf{A}^t$ by $\mathbf{a}_1, \ldots, \mathbf{a}_m$, System 2 has a solution if $\mathbf{c}$ lies in the convex cone generated by $\mathbf{a}_1, \ldots, \mathbf{a}_m$. System 1 has a solution if the closed convex cone $\{\mathbf{x} : \mathbf{Ax} \leq \mathbf{0}\}$ and the open half-space $\{\mathbf{x} : \mathbf{c}^t\mathbf{x} > 0\}$ have a nonempty intersection. These two cases are illustrated geometrically in Figure 2.10.

2.4.5 Theorem (Farkas's Theorem)

Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{c}$ be an n-vector. Then exactly one of the following two systems has a solution:

***System* 1**: $\mathbf{Ax} \leq \mathbf{0}$ and $\mathbf{c}^t\mathbf{x} > 0$ for some $\mathbf{x} \in R^n$.

***System* 2**: $\mathbf{A}^t\mathbf{y} = \mathbf{c}$ and $\mathbf{y} \geq \mathbf{0}$ for some $\mathbf{y} \in R^m$.

Proof

Suppose that System 2 has a solution; that is, there exists $\mathbf{y} \geq \mathbf{0}$ such that $\mathbf{A}^t\mathbf{y} = \mathbf{c}$. Let $\mathbf{x}$ be such that $\mathbf{Ax} \leq \mathbf{0}$. Then $\mathbf{c}^t\mathbf{x} = \mathbf{y}^t\mathbf{Ax} \leq 0$. Hence, System 1 has no solution. Now suppose that System 2 has no solution. Form the set $S = \{\mathbf{x} : \mathbf{x} = \mathbf{A}^t\mathbf{y}, \mathbf{y} \geq \mathbf{0}\}$. Note that S is a closed convex set and that $\mathbf{c} \notin S$. By Theorem

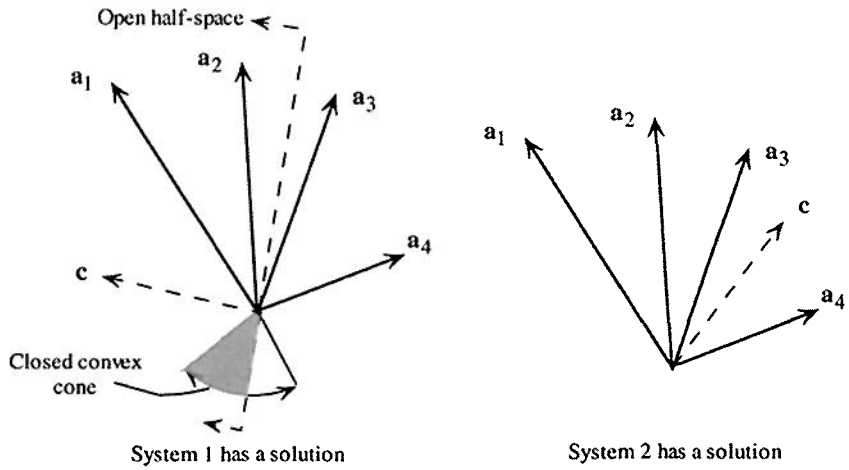


Figure 2.10 Farkas's theorem.

2.4.4, there exists a vector $\mathbf{p} \in R^n$ and a scalar α such that $\mathbf{p}^t\mathbf{c} > \alpha$ and $\mathbf{p}^t\mathbf{x} \leq \alpha$ for all $\mathbf{x} \in S$. Since $\mathbf{0} \in S$, $\alpha \geq 0$, so $\mathbf{p}^t\mathbf{c} > 0$. Also, $\alpha \geq \mathbf{p}^t\mathbf{A}^t\mathbf{y} = \mathbf{y}^t\mathbf{A}\mathbf{p}$ for all $\mathbf{y} \geq \mathbf{0}$. Since $\mathbf{y} \geq \mathbf{0}$ can be made arbitrarily large, the last inequality implies that $\mathbf{A}\mathbf{p} \leq \mathbf{0}$. We have therefore constructed a vector $\mathbf{p} \in R^n$ such that $\mathbf{A}\mathbf{p} \leq \mathbf{0}$ and $\mathbf{c}^t\mathbf{p} > 0$. Hence, System 1 has a solution, and the proof is complete.

Corollary 1 (Gordan's Theorem)

Let $\mathbf{A}$ be an $m \times n$ matrix. Then, exactly one of the following two systems has a solution:

***System* 1**: $\mathbf{A}\mathbf{x} < \mathbf{0}$ for some $\mathbf{x} \in R^n$.

***System* 2**: $\mathbf{A}^t\mathbf{y} = \mathbf{0}$, $\mathbf{y} \geq \mathbf{0}$ for some nonzero $\mathbf{y} \in R^m$.

Proof

Note that System 1 can be written equivalently as $\mathbf{A}\mathbf{x} + \mathbf{e}s \leq \mathbf{0}$ for some $\mathbf{x} \in R^n$ and $s > 0$, $s \in R$, where $\mathbf{e}$ is a vector of m ones. Rewriting this in the form of System 1 of Theorem 2.4.5, we get $[\mathbf{A}\ \mathbf{e}]\begin{bmatrix}\mathbf{x}\\ s\end{bmatrix} \leq \mathbf{0}$, and $(0,\ldots,0,1)\begin{pmatrix}\mathbf{x}\\ s\end{pmatrix} > 0$ for some $\begin{pmatrix}\mathbf{x}\\ s\end{pmatrix} \in R^{n+1}$. By Theorem 2.4.5, the associated System 2 states that $\begin{bmatrix}\mathbf{A}^t\\ \mathbf{e}^t\end{bmatrix}\mathbf{y} = (0,\ldots,0,1)^t$ and $\mathbf{y} \geq \mathbf{0}$ for some $\mathbf{y} \in R^m$; that is, $\mathbf{A}^t\mathbf{y} = \mathbf{0}$, $\mathbf{e}^t\mathbf{y} = 1$, and

$\mathbf{y} \geq \mathbf{0}$ for some $\mathbf{y} \in R^m$. This is equivalent to System 2 of the corollary, and hence the result follows.

Corollary 2

Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{c}$ be an n-vector. Then exactly one of the following two systems has a solution:

***System* 1**: $\mathbf{Ax} \leq \mathbf{0}$, $\mathbf{x} \geq \mathbf{0}$, $\mathbf{c}^t\mathbf{x} > 0$ for some $\mathbf{x} \in R^n$.

***System* 2**: $\mathbf{A}^t\mathbf{y} \geq \mathbf{c}$, $\mathbf{y} \geq \mathbf{0}$ for some $\mathbf{y} \in R^m$.

Proof

The result follows by writing the first set of constraints of System 2 as equalities and, accordingly, replacing $\mathbf{A}^t$ in the theorem by $[\mathbf{A}^t, -\mathbf{I}]$.

Corollary 3

Let $\mathbf{A}$ be an $m \times n$ matrix, $\mathbf{B}$ be an $\ell \times n$ matrix, and $\mathbf{c}$ be an n-vector. Then exactly one of the following two systems has a solution:

***System* 1**: $\mathbf{Ax} \leq \mathbf{0}$, $\mathbf{Bx} = \mathbf{0}$, $\mathbf{c}^t\mathbf{x} > 0$ for some $\mathbf{x} \in R^n$.

***System* 2**: $\mathbf{A}^t\mathbf{y} + \mathbf{B}^t\mathbf{z} = \mathbf{c}$, $\mathbf{y} \geq \mathbf{0}$ for some $\mathbf{y} \in R^m$ and $\mathbf{z} \in R^\ell$.

Proof

The result follows by writing $\mathbf{z} = \mathbf{z}_1 - \mathbf{z}_2$, where $\mathbf{z}_1 \geq \mathbf{0}$ and $\mathbf{z}_2 \geq \mathbf{0}$ in System 2 and, accordingly, replacing $\mathbf{A}^t$ in the theorem by $[\mathbf{A}^t, \mathbf{B}^t, -\mathbf{B}^t]$.

Support of Sets at Boundary Points

2.4.6 Definition

Let S be a nonempty set in R^n, and let $\bar{\mathbf{x}} \in \partial S$. A hyperplane $H = \{\mathbf{x} : \mathbf{p}^t(\mathbf{x} - \bar{\mathbf{x}}) = 0\}$ is called a *supporting hyperplane* of S at $\bar{\mathbf{x}}$ if either $S \subseteq H^+$, that is, $\mathbf{p}^t(\mathbf{x} - \bar{\mathbf{x}}) \geq 0$ for each $\mathbf{x} \in S$, or else, $S \subseteq H^-$, that is, $\mathbf{p}^t(\mathbf{x} - \bar{\mathbf{x}}) \leq 0$ for each $\mathbf{x} \in S$. If, in addition, $S \not\subseteq H$, H is called a *proper supporting hyperplane* of S at $\bar{\mathbf{x}}$.

Note that Definition 2.4.6 can be stated equivalently as follows. The hyperplane $H = \{\mathbf{x} : \mathbf{p}^t(\mathbf{x} - \bar{\mathbf{x}}) = 0\}$ is a supporting hyperplane of S at $\bar{\mathbf{x}} \in \partial S$ if $\mathbf{p}^t\bar{\mathbf{x}} = \inf\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S\}$ or else $\mathbf{p}^t\bar{\mathbf{x}} = \sup\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S\}$. This follows by noting that either $\bar{\mathbf{x}} \in S$, or if $\bar{\mathbf{x}} \notin S$, then since $\bar{\mathbf{x}} \in \partial S$, there exist points in S arbitrarily

close to $\bar{\mathbf{x}}$ and hence arbitrarily close in the value of the function $\mathbf{p}^t\mathbf{x}$ to the value $\mathbf{p}^t\bar{\mathbf{x}}$.

Figure 2.11 shows some examples of supporting hyperplanes. The figure illustrates the cases of a unique supporting hyperplane at a boundary point, an infinite number of supporting hyperplanes at a boundary point, a hyperplane that supports the set at more than one point, and finally, an improper supporting hyperplane that contains the entire set.

We now prove that a convex set has a supporting hyperplane at each boundary point (see Figure 2.12). As a corollary, a result similar to Theorem 2.4.4, where S is not required to be closed, follows.

2.4.7 Theorem

Let S be a nonempty convex set in R^n, and let $\bar{\mathbf{x}} \in \partial S$. Then there exists a hyperplane that supports S at $\bar{\mathbf{x}}$; that is, there exists a nonzero vector $\mathbf{p}$ such that $\mathbf{p}^t(\mathbf{x}-\bar{\mathbf{x}}) \leq 0$ for each $\mathbf{x} \in \text{cl } S$.

Proof

Since $\bar{\mathbf{x}} \in \partial S$, there exists a sequence $\{\mathbf{y}_k\}$ not in cl S such that $\mathbf{y}_k \to \bar{\mathbf{x}}$. By Theorem 2.4.4, corresponding to each $\mathbf{y}_k$ there exists a $\mathbf{p}_k$ with norm 1 such that $\mathbf{p}_k^t\mathbf{y}_k > \mathbf{p}_k^t\mathbf{x}$ for each $\mathbf{x} \in \text{cl } S$. (In Theorem 2.4.4, the normal vector can be normalized by dividing it by its norm, so that $\|\mathbf{p}_k\| = 1$.) Since $\{\mathbf{p}_k\}$ is bounded,

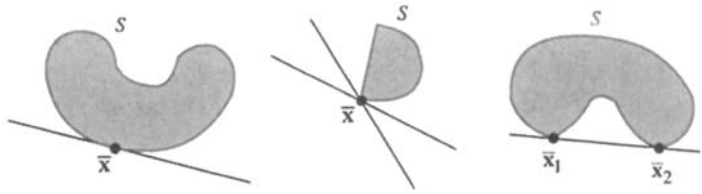


Figure 2.11 Supporting hyperplanes.

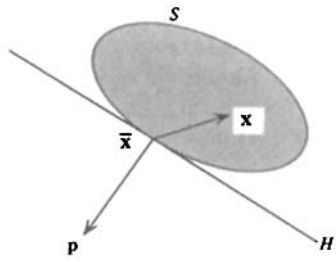


Figure 2.12 Supporting hyperplane.

it has a convergent subsequence $\{\mathbf{p}_k\}_{\mathscr{K}}$ with limit $\mathbf{p}$ whose norm is also equal to 1. Considering this subsequence, we have $\mathbf{p}_k^t \mathbf{y}_k > \mathbf{p}_k^t \mathbf{x}$ for each $\mathbf{x} \in \text{cl } S$. Fixing $\mathbf{x} \in \text{cl } S$ and taking limits as $k \in \mathscr{K}$ approaches ∞, we get $\mathbf{p}^t(\mathbf{x} - \overline{\mathbf{x}}) \leq 0$. Since this is true for each $\mathbf{x} \in \text{cl } S$, the result follows.

Corollary 1

Let S be a nonempty convex set in R^n and $\overline{\mathbf{x}} \notin \text{int } S$. Then there is a nonzero vector $\mathbf{p}$ such that $\mathbf{p}^t(\mathbf{x} - \overline{\mathbf{x}}) \leq 0$ for each $\mathbf{x} \in \text{cl } S$.

Proof

If $\overline{\mathbf{x}} \in \text{cl} S$, the corollary follows from Theorem 2.4.4. On the other hand, if $\overline{\mathbf{x}} \in \partial S$, the corollary reduces to Theorem 2.4.7.

Corollary 2

Let S be a nonempty set in R^n, and let $\mathbf{y} \notin \text{int conv}(S)$. Then there exists a hyperplane that separates S and $\mathbf{y}$.

Proof

The result follows by identifying conv(S) and $\mathbf{y}$ with S and $\mathbf{x}$, respectively, in Corollary 1.

Corollary 3

Let S be a nonempty set in R^n, and let $\overline{\mathbf{x}} \in \partial S \cap \partial$ conv(S). Then there exists a hyperplane that supports S at $\overline{\mathbf{x}}$.

Proof

The result follows by treating conv(S) as the set of Theorem 2.4.7.

Separation of Two Convex Sets

Thus far we have discussed the separation of a convex set and a point not in the set and have also discussed the support of convex sets at boundary points. In addition, if we have two disjoint convex sets, they can be separated by a hyperplane H such that one of the sets belongs to H^+ and the other set belongs to H^-. In fact, this result holds true even if the two sets have some points in common, as long as their interiors are disjoint. This result is made precise by the following theorem.

2.4.8 Theorem

Let S_1 and S_2 be nonempty convex sets in R^n and suppose that $S_1 \cap S_2$ is empty. Then there exists a hyperplane that separates S_1 and S_2; that is, there exists a nonzero vector $\mathbf{p}$ in R^n such that

$$\inf\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S_1\} \geq \sup\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S_2\}.$$

Proof

Let $S = S_1 \ominus S_2 = \{\mathbf{x}_1 - \mathbf{x}_2 : \mathbf{x}_1 \in S_1 \text{ and } \mathbf{x}_2 \in S_2\}$. Note that S is a convex set. Furthermore, $\mathbf{0} \notin S$, because otherwise $S_1 \cap S_2$ will be nonempty. By Corollary 1 of Theorem 2.4.7, there exists a nonzero $\mathbf{p} \in R^n$ such that $\mathbf{p}^t\mathbf{x} \geq 0$ for all $\mathbf{x} \in S$. This means that $\mathbf{p}^t\mathbf{x}_1 \geq \mathbf{p}^t\mathbf{x}_2$ for all $\mathbf{x}_1 \in S_1$ and $\mathbf{x}_2 \in S_2$, and the result follows.

Corollary 1

Let S_1 and S_2 be nonempty convex sets in R^n. Suppose that int S_2 is not empty and that $S_1 \cap \text{int}\, S_2$ is empty. Then there exists a hyperplane that separates S_1 and S_2; that is, there exists a nonzero $\mathbf{p}$ such that

$$\inf\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S_1\} \geq \sup\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S_2\}.$$

Proof

Replace S_2 by int S_2, apply the theorem, and note that

$$\sup\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S_2\} = \sup\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in \text{int}\, S_2\}.$$

Corollary 2

Let S_1 and S_2 be nonempty sets in R^n such that int $\text{conv}(S_i) \neq \varnothing$, for $i = 1, 2$, but int $\text{conv}(S_1) \cap \text{int}\, \text{conv}(S_2) = \varnothing$. Then there exists a hyperplane that separates S_1 and S_2.

Note the importance of assuming nonempty interiors in Corollary 2. Otherwise, for example, two crossing lines in R^2 can be taken as S_1 and S_2 [or as $\text{conv}(S_1)$ and $\text{conv}(S_2)$], and we would have int $\text{conv}(S_1) \cap$ int $\text{conv}(S_2) = \varnothing$. But there does not exist a hyperplane that separates S_1 and S_2.

Gordan's Theorem as a Consequence of Theorem 2.4.8

We shall now prove Gordan's theorem (see Corollary 1 to Theorem 2.4.5) using the existence of a hyperplane that separates two disjoint convex sets. This

theorem is important in deriving optimality conditions for nonlinear programming.

2.4.9 Theorem (Gordan's Theorem)

Let $\mathbf{A}$ be an $m \times n$ matrix. Then exactly one of the following systems has a solution:

***System* 1**: $\mathbf{Ax} < \mathbf{0}$ for some $\mathbf{x} \in R^n$.

***System* 2**: $\mathbf{A}^t\mathbf{p} = \mathbf{0}$ and $\mathbf{p} \geq \mathbf{0}$ for some nonzero $\mathbf{p} \in R^m$.

Proof

We shall first prove that if System 1 has a solution, we cannot have a solution to $\mathbf{A}^t\mathbf{p} = \mathbf{0}$, $\mathbf{p} \geq \mathbf{0}$, $\mathbf{p}$ nonzero. Suppose, on the contrary, that a solution $\hat{\mathbf{p}}$ exists. Then since $\mathbf{A}\hat{\mathbf{x}} < \mathbf{0}$, $\hat{\mathbf{p}} \geq \mathbf{0}$, and $\hat{\mathbf{p}} \neq \mathbf{0}$, we have $\hat{\mathbf{p}}^t\mathbf{A}\hat{\mathbf{x}} < 0$; that is, $\hat{\mathbf{x}}^t\mathbf{A}^t\hat{\mathbf{p}} < 0$. But this contradicts the hypothesis that $\mathbf{A}^t\hat{\mathbf{p}} = \mathbf{0}$. Hence, System 2 cannot have a solution.

Now assume that System 1 has no solution. Consider the following two sets:

$$S_1 = \{\mathbf{z} : \mathbf{z} = \mathbf{Ax}, \mathbf{x} \in R^n\}$$

$$S_2 = \{\mathbf{z} : \mathbf{z} < \mathbf{0}\}.$$

Note that S_1 and S_2 are nonempty convex sets such that $S_1 \cap S_2 = \varnothing$. Then, by Theorem 2.4.8, there exists a hyperplane that separates S_1 and S_2; that is, there exists a nonzero vector $\mathbf{p}$ such that

$$\mathbf{p}^t\mathbf{Ax} \geq \mathbf{p}^t\mathbf{z} \qquad \text{for each } \mathbf{x} \in R^n \text{ and } \mathbf{z} \in \text{cl}S_2.$$

Since each component of $\mathbf{z}$ can be made an arbitrarily large negative number, we must have $\mathbf{p} \geq \mathbf{0}$. Also, by letting $\mathbf{z} = \mathbf{0}$, we must have $\mathbf{p}^t\mathbf{Ax} \geq 0$ for each $\mathbf{x} \in R^n$. By choosing $\mathbf{x} = -\mathbf{A}^t\mathbf{p}$, it follows that $-\left\|\mathbf{A}^t\mathbf{p}\right\|^2 \geq 0$, and thus $\mathbf{A}^t\mathbf{p} = \mathbf{0}$. Hence, System 2 has a solution, and the proof is complete.

Separation Theorem 2.4.8 can be strengthened to avoid trivial separation where both S_1 and S_2 are contained in the separating hyperplane.

2.4.10 Theorem (Strong Separation)

Let S_1 and S_2 be closed convex sets, and suppose that S_1 is bounded. If $S_1 \cap S_2$ is empty, there exists a hyperplane that strongly separates S_1 and S_2; that is, there exists a nonzero $\mathbf{p}$ and $\varepsilon > 0$ such that

$$\inf\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S_1\} \geq \varepsilon + \sup\{\mathbf{p}^t\mathbf{x} : \mathbf{x} \in S_2\}.$$

Proof

Let $S = S_1 \ominus S_2$, and note that S is a convex set and that $\mathbf{0} \notin S$. We shall show that S is closed. Let $\{\mathbf{x}_k\}$ in S converge to $\mathbf{x}$. By the definition of S, $\mathbf{x}_k = \mathbf{y}_k - \mathbf{z}_k$, where $\mathbf{y}_k \in S_1$ and $\mathbf{z}_k \in S_2$. Since S_1 is compact, there is a subsequence $\{\mathbf{y}_k\}_{\mathscr{K}}$ with limit $\mathbf{y}$ in S_1. Since $\mathbf{y}_k - \mathbf{z}_k \to \mathbf{x}$ and $\mathbf{y}_k \to \mathbf{y}$ for $k \in \mathscr{K}$, $\mathbf{z}_k \to \mathbf{z}$ for $k \in \mathscr{K}$. Since S_2 is closed, $\mathbf{z} \in S_2$. Therefore, $\mathbf{x} = \mathbf{y} - \mathbf{z}$ with $\mathbf{y} \in S_1$ and $\mathbf{z} \in S_2$. Therefore, $\mathbf{x} \in S$ and hence S is closed. By Theorem 2.4.4, there is a nonzero $\mathbf{p}$ and an ε such that $\mathbf{p}^t\mathbf{x} \geq \varepsilon$ for each $\mathbf{x} \in S$ and $\mathbf{p}^t\mathbf{0} < \varepsilon$. Therefore, $\varepsilon > 0$. By the definition of S, we conclude that $\mathbf{p}^t\mathbf{x}_1 \geq \varepsilon + \mathbf{p}^t\mathbf{x}_2$ for each $\mathbf{x}_1 \in S_1$ and $\mathbf{x}_2 \in S_2$, and the result follows.

Note the importance of assuming the boundedness of at least one of the sets S_1 and S_2 in Theorem 2.4.10. Figure 2.13 illustrates a situation in R^2 where the boundaries of S_1 and S_2 asymptotically approach the strictly separating hyperplane shown therein. Here S_1 and S_2 are closed convex sets and $S_1 \cap S_2 = \varnothing$, but there does not exist a hyperplane that strongly separates S_1 and S_2. However, if we bound one of the sets, we can obtain a strongly separating hyperplane.

As a direct consequence of Theorem 2.4.10, the following corollary gives a strengthened restatement of the theorem.

Corollary 1

Let S_1 and S_2 be nonempty sets in R^n, and suppose that S_1 is bounded. If cl $\text{conv}(S_1) \cap \text{cl conv}(S_2) = \varnothing$, there exists a hyperplane that strongly separates S_1 and S_2.

2.5 Convex Cones and Polarity

In this section we discuss briefly the notions of convex cones and polar cones. Except for the definition of a (convex) cone, this section may be skipped without loss of continuity.

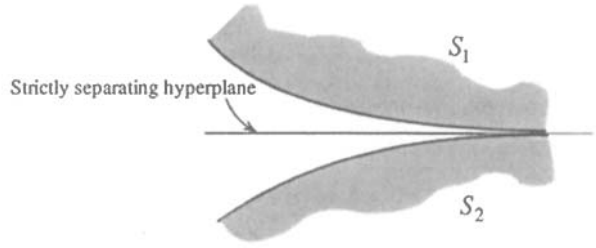


Figure 2.13 Nonexistence of a strongly separating hyperplane.

2.5.1 Definition

A nonempty set C in R^n is called a *cone* with vertex zero if $\mathbf{x} \in C$ implies that $\lambda \mathbf{x} \in C$ for all $\lambda \geq 0$. If, in addition, C is convex, C is called a *convex cone.* Figure 2.14 shows an example of a convex cone and an example of a nonconvex cone.

An important special class of convex cones is that of polar cones, defined below and illustrated in Figure 2.15.

2.5.2 Definition

Let S be a nonempty set in R^n. Then the *polar cone* of S, denoted by S^*, is given by $\{\mathbf{p} : \mathbf{p}^t \mathbf{x} \leq 0 \text{ for all } \mathbf{x} \in S\}$. If S is empty, S^* will be interpreted as R^n.

The following lemma, the proof of which is left as an exercise, summarizes some facts about polar cones.

2.5.3 Lemma

Let S, S_1, and S_2 be nonempty sets in R^n. Then the following statements hold true.

1. S^* is a closed convex cone.
2. $S \subseteq S^{**}$, where S^{**} is the polar cone of S^*.

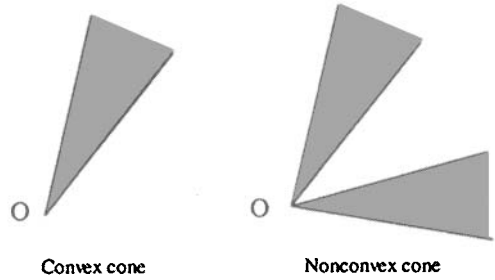


Figure 2.14 Cones.

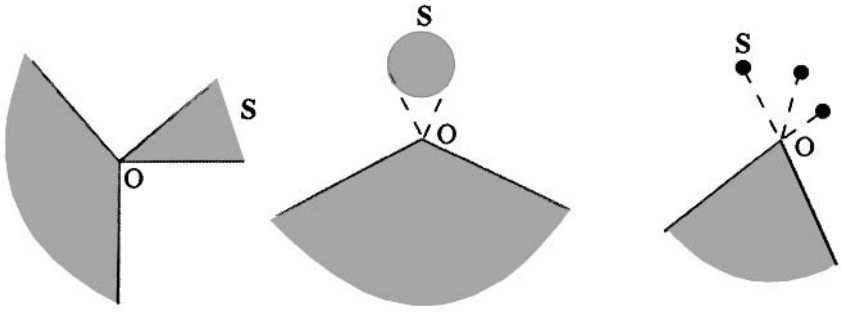


Figure 2.15 Polar cones.

3. $S_1 \subseteq S_2$ implies that $S_2^* \subseteq S_1^*$.

We now prove an important theorem for closed convex cones. As an application of the theorem, we give another derivation of Farkas's theorem.

2.5.4 Theorem

Let C be a nonempty closed convex cone. Then $C = C^{**}$.

Proof

Clearly, $C \subseteq C^{**}$. Now let $\mathbf{x} \in C^{**}$, and suppose, by contradiction, that $\mathbf{x} \notin C$. By Theorem 2.4.4 there exists a nonzero vector $\mathbf{p}$ and a scalar α such that $\mathbf{p}^t\mathbf{y} \leq \alpha$ for all $\mathbf{y} \in C$ and $\mathbf{p}^t\mathbf{x} > \alpha$. But since $\mathbf{y} = \mathbf{0} \in C$, $\alpha \geq 0$, so $\mathbf{p}^t\mathbf{x} > 0$. We now show that $\mathbf{p} \in C^*$. If not, $\mathbf{p}^t\bar{\mathbf{y}} > 0$ for some $\bar{\mathbf{y}} \in C$, and $\mathbf{p}^t(\lambda\bar{\mathbf{y}})$ can be made arbitrarily large by choosing λ arbitrarily large. This contradicts the fact that $\mathbf{p}^t\mathbf{y} \leq \alpha$ for all $\mathbf{y} \in C$. Therefore, $\mathbf{p} \in C^*$. Since $\mathbf{x} \in C^{**} \equiv \{\mathbf{u} : \mathbf{u}^t\mathbf{v} \leq 0$ for all $\mathbf{v} \in C^*\}$, $\mathbf{p}^t\mathbf{x} \leq 0$. This contradicts the fact that $\mathbf{p}^t\mathbf{x} > 0$, and we conclude that $\mathbf{x} \in C$. This completes the proof.

Farkas's Theorem as a Consequence of Theorem 2.5.4

Let $\mathbf{A}$ be an $m \times n$ matrix, and let $C = \{\mathbf{A}^t\mathbf{y} : \mathbf{y} \geq \mathbf{0}\}$. Note that C is a closed convex cone. It can be easily verified that $C^* = \{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{0}\}$. By the theorem, $\mathbf{c} \in C^{**}$ if and only if $\mathbf{c} \in C$. But $\mathbf{c} \in C^{**}$ means that whenever $\mathbf{x} \in C^*$, $\mathbf{c}^t\mathbf{x} \leq 0$, or equivalently, $\mathbf{A}\mathbf{x} \leq \mathbf{0}$ implies that $\mathbf{c}^t\mathbf{x} \leq 0$. By the definition of C, $\mathbf{c} \in C$ means that $\mathbf{c} = \mathbf{A}^t\mathbf{y}$ and $\mathbf{y} \geq \mathbf{0}$. Thus, the result $C = C^{**}$ could be stated as follows: System 1 below is consistent if and only if System 2 has a solution $\mathbf{y}$.

System 1: $\mathbf{A}\mathbf{x} \leq \mathbf{0}$ implies that $\mathbf{c}^t\mathbf{x} \leq 0$.

System 2: $\mathbf{A}^t\mathbf{y} = \mathbf{c}$, $\mathbf{y} \geq \mathbf{0}$.

This statement can be put in the more usual and equivalent form of Farkas's theorem. Exactly one of the following two systems has a solution:

System 1: $\mathbf{A}\mathbf{x} \leq \mathbf{0}$, $\mathbf{c}^t\mathbf{x} > 0$ (i.e., $\mathbf{c} \notin C^{**} = C$).

System 2: $\mathbf{A}^t\mathbf{y} = \mathbf{c}$, $\mathbf{y} \geq \mathbf{0}$ (i.e., $\mathbf{c} \in C$).

2.6 Polyhedral Sets, Extreme Points, and Extreme Directions

In this section we introduce the notions of extreme points and extreme directions for convex sets. We then discuss in more detail their use for the special important case of polyhedral sets.

Polyhedral Sets

Polyhedral sets represent an important special case of convex sets. We have seen from the corollary to Theorem 2.4.4 that any closed convex set is the intersection of all closed half-spaces containing it. In the case of polyhedral sets, only a finite number of half-spaces are needed to represent the set.

2.6.1 Definition

A set S in R^n is called a *polyhedral set* if it is the intersection of a finite number of closed half-spaces; that is, $S = \{\mathbf{x} : \mathbf{p}_i^t\mathbf{x} \leq \alpha_i \text{ for } i = 1,..., m\}$, where $\mathbf{p}_i$ is a nonzero vector and α_i is a scalar for $i = 1,..., m$.

Note that a polyhedral set is a closed convex set. Since an equation can be represented by two inequalities, a polyhedral set can be represented by a finite number of inequalities and/or equations. The following are some typical examples of polyhedral sets, where $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b}$ is an m-vector:

$$S = \{\mathbf{x} : \mathbf{Ax} \leq \mathbf{b}\}$$

$$S = \{\mathbf{x} : \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$$

$$S = \{\mathbf{x} : \mathbf{Ax} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}.$$

Figure 2.16 illustrates the polyhedral set

$$S = \{(x_1, x_2) : -x_1 + x_2 \leq 2,\ x_2 \leq 4,\ x_1 \geq 0,\ x_2 \geq 0\}.$$

Extreme Points and Extreme Directions

We now introduce the concepts of extreme points and extreme directions for convex sets. We then give their full characterizations in the case of polyhedral sets.

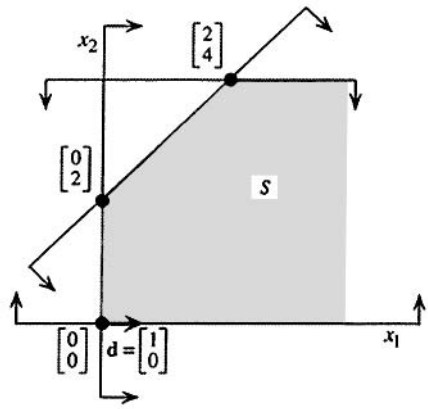


Figure 2.16 Polyhedral set.

2.6.2 Definition

Let S be a nonempty convex set in R^n. A vector $\mathbf{x} \in S$ is called an *extreme point* of S if $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ with $\mathbf{x}_1, \mathbf{x}_2 \in S$, and $\lambda \in (0, 1)$ implies that $\mathbf{x} = \mathbf{x}_1 = \mathbf{x}_2$.

The following are some examples of extreme points of convex sets. We denote the set of extreme points by E and illustrate them in Figure 2.17 by dark points or dark lines as indicated.

1. $S = \{(x_1, x_2) : x_1^2 + x_2^2 \le 1\}$;
   $E = \{(x_1, x_2) : x_1^2 + x_2^2 = 1\}$.
2. $S = \{(x_1, x_2) : x_1 + x_2 \le 2,\ -x_1 + 2x_2 \le 2,\ x_1, x_2 \ge 0\}$;
   $E = \{(0,0)^t, (0,1)^t, (2/3, 4/3)^t, (2,0)^t\}$.
3. S is the polytope generated by $(0,0)^t$, $(1,1)^t$, $(1,3)^t$, $(-2,4)^t$ and $(0,2)^t$; $E = \{(0,0)^t, (1,1)^t, (1,3)^t, (-2,4)^t\}$.

From Figure 2.17 we see that any point of the convex set S can be represented as a convex combination of the extreme points. This turns out to be true for compact convex sets. However, for unbounded sets, we may not be able to represent every point in the set as a convex combination of its extreme points. To illustrate, let $S = \{(x_1, x_2) : x_2 \ge |x_1|\}$. Note that S is convex and closed. However, S contains only one extreme point, the origin, and obviously S is not equal to the collection of convex combinations of its extreme points. To deal with unbounded sets, the notion of extreme directions is needed.

2.6.3 Definition

Let S be a nonempty, closed convex set in R^n. A nonzero vector $\mathbf{d}$ in R^n is called a *direction*, or a *recession direction*, of S if for each $\mathbf{x} \in S$, $\mathbf{x} + \lambda\mathbf{d} \in S$ for all $\lambda \ge 0$. Two directions $\mathbf{d}_1$ and $\mathbf{d}_2$ of S are called *distinct* if $\mathbf{d}_1 \ne \alpha\mathbf{d}_2$ for any $\alpha > 0$. A

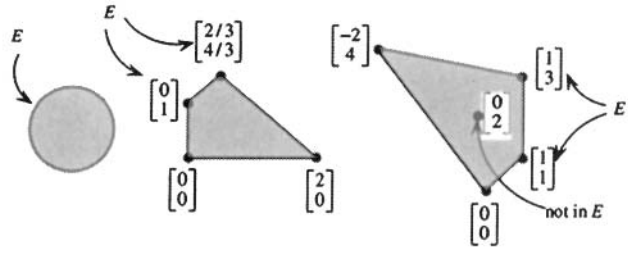


Figure 2.17 Extreme points.

direction $\mathbf{d}$ of S is called an *extreme direction* if it cannot be written as a positive linear combination of two distinct directions; that is, if $\mathbf{d} = \lambda_1\mathbf{d}_1 + \lambda_2\mathbf{d}_2$ for λ_1, $\lambda_2 > 0$, then $\mathbf{d}_1 = \alpha\mathbf{d}_2$ for some $\alpha > 0$.

To illustrate, consider $S = \{(x_1, x_2) : x_2 \geq |x_1|\}$, shown in Figure 2.18. The directions of S are nonzero vectors that make an angle less than or equal to 45° with the vector $(0,1)^t$. In particular, $\mathbf{d}_1 = (1,1)^t$ and $\mathbf{d}_2 = (-1,1)^t$ are two extreme directions of S. Any other direction of S can be represented as a positive linear combination of $\mathbf{d}_1$ and $\mathbf{d}_2$.

Characterization of Extreme Points and Extreme Directions for Polyhedral Sets

Consider the polyhedral set $S = \{\mathbf{x} : \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$, where $\mathbf{A}$ is an $m \times n$ matrix and $\mathbf{b}$ is an m-vector. We assume that the rank of $\mathbf{A}$ is m. If not, then assuming that $\mathbf{Ax} = \mathbf{b}$ is consistent, we can throw away any redundant equations to obtain a full row rank matrix.

Extreme Points Rearrange the columns of $\mathbf{A}$ so that $\mathbf{A} = [\mathbf{B}, \mathbf{N}]$, where $\mathbf{B}$ is an $m \times m$ matrix of full rank and $\mathbf{N}$ is an $m \times (n - m)$ matrix. Let $\mathbf{x}_B$ and $\mathbf{x}_N$ be the vectors corresponding to $\mathbf{B}$ and $\mathbf{N}$, respectively. Then $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$ can be rewritten as follows:

$$\mathbf{Bx}_B + \mathbf{Nx}_N = \mathbf{b} \qquad \text{and} \qquad \mathbf{x}_B \geq \mathbf{0}, \mathbf{x}_N \geq \mathbf{0}.$$

The following theorem gives a necessary and sufficient characterization of an extreme point of S.

2.6.4 Theorem (Characterization of Extreme Points)

Let $S = \{\mathbf{x} : \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$, where $\mathbf{A}$ is an $m \times n$ matrix of rank m and $\mathbf{b}$ is an m-vector. A point $\mathbf{x}$ is an extreme point of S if and only if $\mathbf{A}$ can be decomposed into $[\mathbf{B}, \mathbf{N}]$ such that

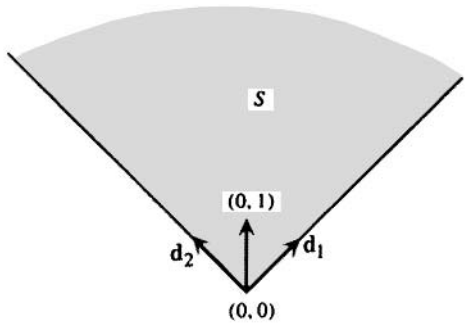


Figure 2.18 Extreme directions.

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_B \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{bmatrix},$$

where $\mathbf{B}$ is an $m \times m$ invertible matrix satisfying $\mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$. Any such solution is called a *basic feasible solution* (BFS) for S.

Proof

Suppose that $\mathbf{A}$ can be decomposed into $[\mathbf{B}, \mathbf{N}]$ with $\mathbf{x} = \begin{bmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{bmatrix}$ and $\mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$. It is obvious that $\mathbf{x} \in S$. Now suppose that $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$ with $\mathbf{x}_1, \mathbf{x}_2 \in S$ for some $\lambda \in (0,1)$. In particular, let $\mathbf{x}_1^t = (\mathbf{x}_{11}^t, \mathbf{x}_{12}^t)$ and $\mathbf{x}_2^t = (\mathbf{x}_{21}^t, \mathbf{x}_{22}^t)$. Then

$$\begin{bmatrix} \mathbf{B}^{-1}\mathbf{b} \\ \mathbf{0} \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{x}_{11} \\ \mathbf{x}_{12} \end{bmatrix} + (1-\lambda) \begin{bmatrix} \mathbf{x}_{21} \\ \mathbf{x}_{22} \end{bmatrix}.$$

Since $\mathbf{x}_{12}, \mathbf{x}_{22} \geq \mathbf{0}$ and $\lambda \in (0, 1)$, it follows that $\mathbf{x}_{12} = \mathbf{x}_{22} = \mathbf{0}$. But this implies that $\mathbf{x}_{11} = \mathbf{x}_{21} = \mathbf{B}^{-1}\mathbf{b}$ and hence $\mathbf{x} = \mathbf{x}_1 = \mathbf{x}_2$. This shows that $\mathbf{x}$ is an extreme point of S. Conversely, suppose that $\mathbf{x}$ is an extreme point of S. Without loss of generality, suppose that $\mathbf{x} = (x_1, ..., x_k, 0, ..., 0)^t$, where $x_1, ..., x_k$ are positive. We shall first show that $\mathbf{a}_1, ..., \mathbf{a}_k$ are linearly independent. By contradiction, suppose that there exist scalars $\lambda_1, ..., \lambda_k$ not all zero such that $\sum_{j=1}^{k} \lambda_j \mathbf{a}_j = \mathbf{0}$. Let $\lambda = (\lambda_1, ..., \lambda_k, 0, ..., 0)^t$. Construct the following two vectors, where $\alpha > 0$ is chosen such that $\mathbf{x}_1, \mathbf{x}_2 \geq \mathbf{0}$:

$$\mathbf{x}_1 = \mathbf{x} + \alpha\lambda \qquad \text{and} \qquad \mathbf{x}_2 = \mathbf{x} - \alpha\lambda.$$

Note that

$$\mathbf{A}\mathbf{x}_1 = \mathbf{A}\mathbf{x} + \alpha\mathbf{A}\lambda = \mathbf{A}\mathbf{x} + \alpha \sum_{j=1}^{k} \lambda_j \mathbf{a}_j = \mathbf{b},$$

and similarly $\mathbf{A}\mathbf{x}_2 = \mathbf{b}$. Therefore, $\mathbf{x}_1, \mathbf{x}_2 \in S$, and since $\alpha > 0$ and $\lambda \neq \mathbf{0}$, $\mathbf{x}_1$ and $\mathbf{x}_2$ are distinct. Moreover, $\mathbf{x} = (1/2)\mathbf{x}_1 + (1/2)\mathbf{x}_2$. This contradicts the fact that $\mathbf{x}$ is an extreme point. Thus, $\mathbf{a}_1, ..., \mathbf{a}_k$ are linearly independent, and since $\mathbf{A}$ has rank m, $m - k$ of the last $n - k$ columns may be chosen such that they, together with the first k columns, form a linearly independent set of m-vectors. To simplify the notation, suppose that these columns are $\mathbf{a}_{k+1}, ..., \mathbf{a}_m$. Thus, $\mathbf{A}$ can be written as

$\mathbf{A} = [\mathbf{B}, \mathbf{N}]$, where $\mathbf{B} = [\mathbf{a}_1,...,\mathbf{a}_m]$ is of full rank. Furthermore, $\mathbf{B}^{-1}\mathbf{b} = (x_1,...,x_k, 0,...,0)^t$, and since $x_j > 0$ for $j = 1,..., k$, $\mathbf{B}^{-1}\mathbf{b} \geq \mathbf{0}$. This completes the proof.

Corollary

The number of extreme points of S is finite.

Proof

The number of extreme points is less than or equal to

$$\binom{n}{m} = \frac{n!}{m!(n-m)!},$$

which is the maximum number of possible ways to choose m columns of $\mathbf{A}$ to form $\mathbf{B}$.

Whereas the above corollary proves that a polyhedral set of the form $\{\mathbf{x} : \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ has a finite number of extreme points, the following theorem shows that every nonempty polyhedral set of this form must have at least one extreme point.

2.6.5 Theorem (Existence of Extreme Points)

Let $S = \{\mathbf{x} : \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ be nonempty, where $\mathbf{A}$ is an $m \times n$ matrix of rank m and $\mathbf{b}$ is an m-vector. Then S has at least one extreme point.

Proof

Let $\mathbf{x} \in S$ and, without loss of generality, suppose that $\mathbf{x} = (x_1,...,x_k, 0,...,0)^t$, where $x_j > 0$ for $j = 1,..., k$. If $\mathbf{a}_1,...,\mathbf{a}_k$ are linearly independent, $k \leq m$ and $\mathbf{x}$ is an extreme point. Otherwise, there exist scalars $\lambda_1,...,\lambda_k$ with at least one positive component such that $\sum_{j=1}^{k} \lambda_j \mathbf{a}_j = \mathbf{0}$. Define $\alpha > 0$ as follows:

$$\alpha = \underset{1 \leq j \leq k}{\text{minimum}} \left\{ \frac{x_j}{\lambda_j} : \lambda_j > 0 \right\} = \frac{x_i}{\lambda_i}, \text{ say.}$$

Consider the point $\mathbf{x}'$ whose jth component x'_j is given by

$$x'_j = \begin{cases} x_j - \alpha\lambda_j & \text{for } j = 1,...,k \\ 0 & \text{for } j = k+1,...,n. \end{cases}$$

Note that $x'_j \geq 0$ for $j = 1,..., k$ and $x'_j = 0$ for $j = k + 1,..., n$. Moreover, $x'_i = 0$, and

$$\sum_{j=1}^{n} \mathbf{a}_j x'_j = \sum_{j=1}^{k} \mathbf{a}_j (x_j - \alpha\lambda_j) = \sum_{j=1}^{k} \mathbf{a}_j x_j - \alpha \sum_{j=1}^{k} \mathbf{a}_j \lambda_j = \mathbf{b} - \mathbf{0} = \mathbf{b}.$$

Thus, so far, we have constructed a new point $\mathbf{x}'$ with at most $k - 1$ positive components. The process is continued until the positive components correspond to linearly independent columns, which results in an extreme point. Thus, we have shown that S has at least one extreme point, and the proof is complete.

Extreme Directions Let $S = \{\mathbf{x} : \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \neq \varnothing$, where $\mathbf{A}$ is an $m \times n$ matrix of rank m. By definition, a nonzero vector $\mathbf{d}$ is a direction of S if $\mathbf{x} + \lambda\mathbf{d} \in S$ for each $\mathbf{x} \in S$ and each $\lambda \geq 0$. Noting the structure of S, it is clear that $\mathbf{d} \neq \mathbf{0}$ is a direction of S if and only if

$$\mathbf{Ad} = \mathbf{0}, \qquad \mathbf{d} \geq \mathbf{0}.$$

In particular, we are interested in the characterization of extreme directions of S.

2.6.6 Theorem (Characterization of Extreme Directions)

Let $S = \{\mathbf{x} : \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \neq \varnothing$, where $\mathbf{A}$ is an $m \times n$ matrix of rank m and $\mathbf{b}$ is an m-vector. A vector $\bar{\mathbf{d}}$ is an extreme direction of S if and only if $\mathbf{A}$ can be decomposed into $[\mathbf{B}, \mathbf{N}]$ such that $\mathbf{B}^{-1}\mathbf{a}_j \leq \mathbf{0}$ for some column $\mathbf{a}_j$ of $\mathbf{N}$, and $\bar{\mathbf{d}}$ is a positive multiple of $\mathbf{d} = \begin{pmatrix} -\mathbf{B}^{-1}\mathbf{a}_j \\ \mathbf{e}_j \end{pmatrix}$, where $\mathbf{e}_j$ is an $n - m$ vector of zeros except for a 1 in position j.

Proof

If $\mathbf{B}^{-1}\mathbf{a}_j \leq \mathbf{0}$, $\mathbf{d} \geq \mathbf{0}$. Furthermore, $\mathbf{Ad} = \mathbf{0}$, so that $\mathbf{d}$ is a direction of S. We now show that $\mathbf{d}$ is indeed an extreme direction. Suppose that $\mathbf{d} = \lambda_1\mathbf{d}_1 + \lambda_2\mathbf{d}_2$, where $\lambda_1, \lambda_2 > 0$ and $\mathbf{d}_1, \mathbf{d}_2$ are directions of S. Noting that $n - m - 1$ components of $\mathbf{d}$ are equal to zero, the corresponding components of $\mathbf{d}_1$ and $\mathbf{d}_2$ must also be equal to zero. Thus, $\mathbf{d}_1$ and $\mathbf{d}_2$ could be written as follows:

$$\mathbf{d}_1 = \alpha_1 \begin{pmatrix} \mathbf{d}_{11} \\ \mathbf{e}_j \end{pmatrix}, \qquad \mathbf{d}_2 = \alpha_2 \begin{pmatrix} \mathbf{d}_{21} \\ \mathbf{e}_j \end{pmatrix},$$

where $\alpha_1, \alpha_2 > 0$. Noting that $\mathbf{Ad}_1 = \mathbf{Ad}_2 = \mathbf{0}$, it can easily be verified that $\mathbf{d}_{11} = \mathbf{d}_{21} = -\mathbf{B}^{-1}\mathbf{a}_j$. Thus, $\mathbf{d}_1$ and $\mathbf{d}_2$ are not distinct, which implies that $\mathbf{d}$ is an extreme direction. Since $\bar{\mathbf{d}}$ is a positive multiple of $\mathbf{d}$, it is also an extreme direction.

Conversely, suppose that $\bar{\mathbf{d}}$ is an extreme direction of S. Without loss of generality, suppose that