Nonlinear Programming

Theory and Algorithms

Board of Advisors, Engineering

A. H-S. Ang University of Illinois	Civil Engineering — Systems and Probability
Donald S. Berry Northwestern University	Transportation Engineering
James Gere Stanford University	Civil Engineering and Applied Mechanics
J. Stuart Hunter Princeton University	Engineering Statistics
T. William Lambe *R. V. Whitman* Massachusetts Institute of Technology	Civil Engineering — Soil Mechanics
Perry L. McCarty Stanford University	Environmental Engineering
Don T. Phillips Texas A & M University	Industrial Engineering
Dale Rudd University of Wisconsin	Chemical Engineering
Robert F. Steidel, Jr. University of California — Berkeley	Mechanical Engineering
R. N. White Cornell University	Civil Engineering — Structures

Nonlinear Programming

Theory and Algorithms

Mokhtar S. Bazaraa
C. M. Shetty
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia

John Wiley & Sons
New York Chichester Brisbane Toronto

Copyright © 1979, by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Sections 107 and 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission or further information should be addressed to the Permissions Department, John Wiley & Sons.

Library of Congress Cataloging in Publication Data

Bazaraa, M S 1943–
Nonlinear programming.

Bibliography: p.
Includes indexes.
1. Nonlinear programming. I. Shetty, C. M., 1929– joint author. II. Title.
T57.8.B39 519.7'6 78-986
ISBN 0-471-78610-1
Printed in the United States of America
10 9 8 7 6 5 4 3

Dedicated to
OUR PARENTS

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 the simplex method for linear programming and the advent of high-speed computers have made linear programming an important tool for solving problems in diverse fields. However, many realistic problems cannot be adequately represented as a linear program owing to the nonlinearity of the objective function and/or the nonlinearity of any of the constraints. Efforts to solve nonlinear problems efficiently have made rapid progress during the past two 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 the computational method itself.

In preparing this book, a special effort has been made to see 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 and inventory control and from highway design. Examples of a two-bar truss design and a two-bearing journal design are given. Steady-state conditions of an electric 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. Finally, nonlinear models arising in stochastic programming and in location theory are discussed.

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. Generalization of convex functions and their interrelationships are also discussed, since nonlinear programming algorithms suitable for convex functions can be used for a more general class involving pseudoconvex and quasiconvex functions.

Part 2, which includes Chapters 4, 5, and 6, covers optimality conditions and duality. In Chapter 4, the classical Fritz John and Kuhn–Tucker optimality conditions are developed both for inequality- and equality-constrained problems. The topic of constraint qualifications is introduced in Chapter 5. Chapter 6 deals with Lagrangian duality and saddle point optimality conditions. Duality theorems, properties of the dual function, and methods for solving the dual problem are discussed. In addition to Lagrangian duality, there are several other duality formulations in nonlinear programming, such as conjugate duality, min-max duality, surrogate duality, and symmetric duality. Among these duality formulations, the Lagrangian duality seems to be the most promising in the area of algorithmic development. In addition, the results that could be obtained from each of the duality formulations are comparable. In view of these factors and of the space limitation, we have elected to discuss the Lagrangian duality in the text and only introduce other duality formulations in the exercises.

Part 3, consisting of Chapters 7 through 11, discusses 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 to prove the convergence of methods developed later in the book. A brief discussion is also given on criteria that could be used to evaluate algorithms. Chapter 8 deals with optimization of a function without constraints. In particular, we discuss several methods for performing line search as well as methods for minimizing a function of several variables. Methods using both derivative and derivative-free information are

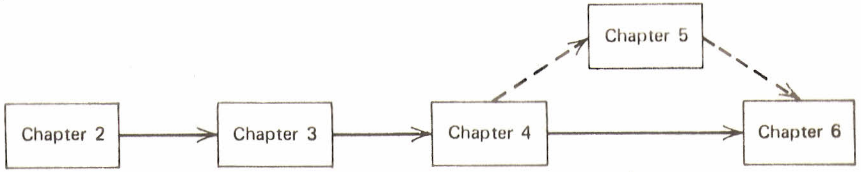
discussed. Methods that are based on the concept of conjugacy are also covered. In Chapters 8 through 11 we prove convergence of each of the algorithms presented. The topic of the order of convergence, briefly introduced in Chapter 7, is not pursued further in view of space limitation and the level of the text. Chapter 9 discusses penalty and barrier function methods for solving nonlinear programs in which essentially the problem is converted into a sequence of unconstrained problems. Chapter 10 discusses 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–Veinott to assure convergence are discussed. The gradient projection method of Rosen, the reduced gradient of Wolfe, and the convex simplex method of Zangwill are also important implementations of the concept of the method of feasible directions and are all discussed in Chapter 10. Chapter 11 deals with some special problems of linear constraints that can be solved by the simplex method for linear programs with minor modifications. In particular, we discuss quadratic, separable, and linear fractional programming. Quadratic programs are solved using Lemke's complementary pivoting algorithm introduced early in the chapter.

This book can be used both as a reference for topics on 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 frequently used in the book.

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

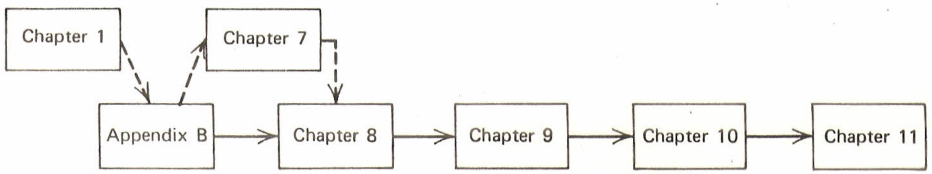
1. *Foundations of Optimization*

This course is meant for undergraduate students in applied mathematics and 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 on constraint qualifications could be omitted without loss of continuity. A reader familiar with linear programming may also skip Section 2.6.



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 analysis 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.



We thank Dr. Robert N. Lehrer, director of the School of Industrial and Systems Engineering at the Georgia Institute of Technology, for his support in the preparation of this manuscript. We have also conferred actively with Dr. Jamie J. Goode of the School of Mathematics, Georgia Institute of Technology, on several occasions. Example 7.3.3 on closedness of composite maps is credited to him. We are deeply indebted for his friendship and active cooperation. Finally, we thank Mrs. Carolyn Piersma, Mrs. Joene Owen, and Ms. Kaye Watkins who successfully managed to decipher and type the several drafts of the manuscript, which sometimes looked as if it were written in Arabic or Hindi.

Atlanta, Georgia
January 1, 1979

Mokhtar S. Bazaraa
C. M. Shetty

Contents

Part 2 Optimality Conditions and Duality

Part 3 Algorithms and Their Convergence

Nonlinear Programming

Theory and Algorithms

Chapter 1

Introduction

Engineers and operations researchers are traditionally involved in problem solving. The problem may involve arriving at an optimal design, allocating scarce resources, 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? During the past three decades, in response to these pressures there has been a very rapid growth of optimization models and techniques. Fortunately, the growth of large and fast 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 out hypotheses, have significantly aided 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 analysis. By that time, efficient procedures and computer codes were available to handle such problems. This book is concerned with nonlinear programming, including the characterization of optimal solutions and the development of computational 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, and the discussion is not intended to be exhaustive.

1.1 Problem Statement and Basic Definitions

Consider the following *nonlinear programming problem:*

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

where $f, g_1, \ldots, g_m, h_1, \ldots, h_l$ are functions defined on E_n, X is a subset of E_n, and $\mathbf{x}$ is a vector of n components $x_1, \ldots, x_n$. The above problem must be solved for the values of the variables $x_1, \ldots, x_n$ 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 $g_i(\mathbf{x}) \le 0$ for $i = 1, \ldots, m$ is called an *inequality constraint*, and each of the constraints $h_i(\mathbf{x}) = 0$ for $i = 1, \ldots, l$ is called an *equality constraint.* 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 $f(\mathbf{x}) \ge f(\bar{\mathbf{x}})$ for each feasible point $\mathbf{x}$. Such a point $\bar{\mathbf{x}}$ is called an *optimal solution*, or simply a *solution*, to the problem.

Needless to say, a nonlinear programming problem can be stated as a maximization problem, and the inequality constraints can be written in the form $g_i(\mathbf{x}) \ge 0$ for $i = 1, \ldots, m$. 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:

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

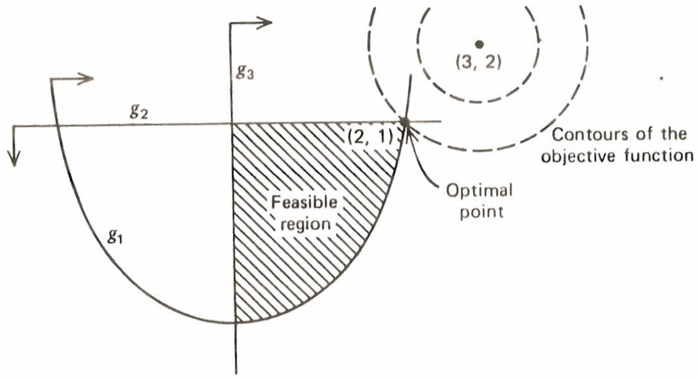


Figure 1.1 Geometric solution of a nonlinear problem.

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 the point in the feasible region with the smallest possible $(x_1-3)^2 + (x_2-2)^2$. 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 value c. Since we wish to minimize c, we must find the circle with 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 with 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 with more than two variables or those with complicated objective and constraint functions.

Notation

The following notation will be used throughout the book. Vectors are denoted by boldface lowercase Roman letters, such as $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. All vectors are column vectors unless explicitly stated otherwise. Row vectors are the transpose of column vectors; for example, $\mathbf{x}^t$ denotes the row vector $(x_1, \ldots, x_n)$.

The n-dimensional *Euclidian space*, the collection of all vectors of dimension n, is denoted by E_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 $\mathbf{g}$ and $\boldsymbol{\psi}$. 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 Some 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 will 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:

$$\mathbf{y}_k = \mathbf{y}_{k-1} + \boldsymbol{\phi}_k(\mathbf{y}_{k-1}, \mathbf{u}_k)$$

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

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

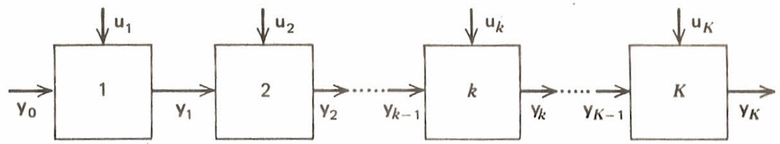


Figure 1.2 Illustration of a discrete control system.

restrictions:

$$\mathbf{y}_k \in Y_k \qquad \text{for } k = 1, \ldots, K$$
$$\mathbf{u}_k \in U_k \qquad \text{for } k = 1, \ldots, K$$
$$\boldsymbol{\psi}(\mathbf{y}_0, \ldots, \mathbf{y}_K, \mathbf{u}_1, \ldots, \mathbf{u}_K) \in D$$

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

$$\begin{array}{lll} \text{Minimize} & \alpha(\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_K, \mathbf{u}_1, \ldots, \mathbf{u}_K) & \\ \text{subject to} & \mathbf{y}_k = \mathbf{y}_{k-1} + \boldsymbol{\phi}_k(\mathbf{y}_{k-1}, \mathbf{u}_k) & \text{for } k = 1, \ldots, K \\ & \mathbf{y}_k \in Y_k & \text{for } k = 1, \ldots, K \\ & \mathbf{u}_k \in U_k & \text{for } k = 1, \ldots, K \\ & \boldsymbol{\psi}(\mathbf{y}_0, \ldots, \mathbf{y}_K, \mathbf{u}_1, \ldots, \mathbf{u}_K) \in D & \end{array}$$

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 be easily verified that the above problem can be stated as the nonlinear programming problem introduced in Section 1.1.

A 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 fired 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 $1, \ldots, K$ 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:

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

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 shall 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:

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

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

$$\mathbf{y} \in Y$$
$$\mathbf{u} \in U$$
$$\boldsymbol{\psi}(\mathbf{y}, \mathbf{u}) \in D$$

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:

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

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 such that $K\Delta = T$. Denoting $\mathbf{y}(k\Delta)$ by $\mathbf{y}_k$ and $\mathbf{u}(k\Delta)$ by $\mathbf{u}_k$, the above problem can be approximated as follows, where the initial state $\mathbf{y}_0$ is given.

$$\begin{array}{lll} \text{Minimize} & \sum_{k=1}^{K} \alpha(\mathbf{y}_k, \mathbf{u}_k) & \\ \text{subject to} & \mathbf{y}_k = \mathbf{y}_{k-1} + \boldsymbol{\phi}_k(\mathbf{y}_{k-1}, \mathbf{u}_k) & \text{for } k=1,\ldots,K \\ & \mathbf{y}_k \in Y_k & \text{for } k=1,\ldots,K \\ & \mathbf{u}_k \in U_k & \text{for } k=1,\ldots,K \\ & \boldsymbol{\psi}(\mathbf{y}_0, \ldots, \mathbf{y}_K, \mathbf{u}_1, \ldots, \mathbf{u}_K) \in D & \end{array}$$

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

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

where $\ddot{y}(t)$ is the acceleration at time t, and g is the deceleration due to gravity. Further 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:

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

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

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

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 1. Denoting the force, altitude, and velocity at the end of period k by u_k, $y_{1,k}$, and $y_{2,k}$ respectively, the above problem can be approximated by the following nonlinear program, where $y_{1,0}=y_{2,0}=0$.

$$\begin{array}{lll} \text{Minimize} & \displaystyle\sum_{k=1}^{K} |u_k| & \\ \text{subject to} & y_{1,k}-y_{1,k-1}=y_{2,k-1} & \text{for } k=1,\ldots,K \\ & y_{2,k}-y_{2,k-1}=u_k-mg & \text{for } k=1,\ldots,K \\ & |u_k|\le b & \text{for } k=1,\ldots,K \\ & y_{1,K}=\bar{y} & \end{array}$$

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

An Example of Highway Construction Suppose that a road is to be constructed over an 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 $t\in[0, T]$. Find the equation describing the height of the road $y(t)$ for $t\in[0, T]$.

In order to avoid excessive slopes on the road, the maximum slope must not exceed b_1, 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, 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:

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

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 1. Denoting $c(k)$, $y_1(k)$, and $y_2(k)$ by c_k, $y_{1,k}$, and $y_{2,k}$, the above problem can be approximated by the following nonlinear program:

$$\begin{array}{lll} \text{Minimize} & \displaystyle\sum_{k=1}^{K} |y_{1,k}-c_k| & \\ \text{subject to} & y_{1,k}-y_{1,k-1}=y_{2,k-1} & \text{for } k=1,\ldots,K \\ & -b_1\le y_{2,k}\le b_1 & \text{for } k=0,\ldots,K-1 \\ & -b_2\le y_{2,k}-y_{2,k-1}\le b_2 & \text{for } k=1,\ldots,K-1 \\ & y_{1,0}=a & \\ & y_{1,K}=b & \end{array}$$

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

B. Structural Design

Structural designers have traditionally endeavoured 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 which 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 $2W$ and so that the total weight of the truss will be minimized.

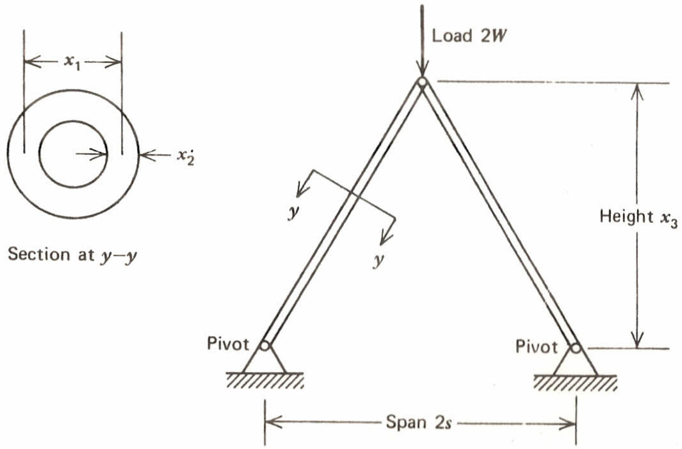


Figure 1.3 Example of a two-bar truss.

Denote the average tube diameter, the 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:

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

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], Johnson [1971], and Fox [1971] give several examples of optimum mechanical design 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 optimum design for a bearing journal.

A 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.

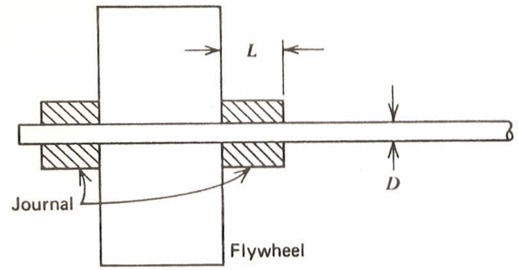


Figure 1.4 Journal bearing assembly.

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 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 $\hat{h}_0$ = minimum oil film thickness to prevent metal-to-metal contact, and
δ = 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, the nature of the torque impulses, and 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

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

where k_3 is a constant depending upon 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 $2c \geq W$ 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:

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

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

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

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

$$0 \leq \left(1 - \frac{h_0}{\delta}\right) \leq \hat{e}$$

$$D \geq 0$$

$$L \geq 0$$

For a thorough discussion of this problem, the reader may refer to Asimov [1962]. The reader can also form the model to minimize the twist angle subject to the frictional moment being within a given maximum limit M'. One could also conceive of an objective function involving both the frictional

moment and the angle of twist, if a proper weight for these factors is 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 was perhaps the first to investigate the relationship between electrical circuit theory, mathematical programming, and duality. The following discussion is based on the pioneering work of Dennis [1959].

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 each branch and the nodes it connects 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 only in one direction and consumes zero power regardless of the branch current or voltage. Denoting the latter by v_d, this can be stated as

$$c_d \geq 0 \qquad v_d \geq 0 \qquad v_d c_d = 0 \tag{1.1}$$

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

$$v_r = -r c_r \tag{1.2}$$

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

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

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,

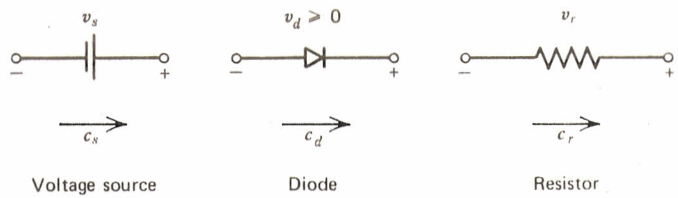


Figure 1.5 Typical electrical devices in a circuit.

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 with a number of branches can be described by a *node-branch incidence* matrix $\mathbf{N}$ whose rows correspond to the nodes and whose columns correspond to the branches. A typical element n_{ij} of $\mathbf{N}$ 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 with several voltage sources, diodes, and resistors, let $\mathbf{N}_S$ denote the node-branch incidence matrix for all the branches with voltage sources, $\mathbf{N}_D$ denote the node-branch incidence matrix for all branches with diodes, and $\mathbf{N}_R$ denote the node-branch incidence matrix for all branches with resistors. Then, without loss of generality, we can partition $\mathbf{N}$ as

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

Likewise, the column vector $\mathbf{c}$, representing the branch currents, can be partitioned as

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

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

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

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) through (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.

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

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 be readily verified that, letting $\mathbf{v}_D = \mathbf{u}_0$, $\mathbf{p} = \mathbf{u}$ and noting (1.7), the above conditions are precisely the equilibrium conditions (1.4) through (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.

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

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) through (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,

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

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.

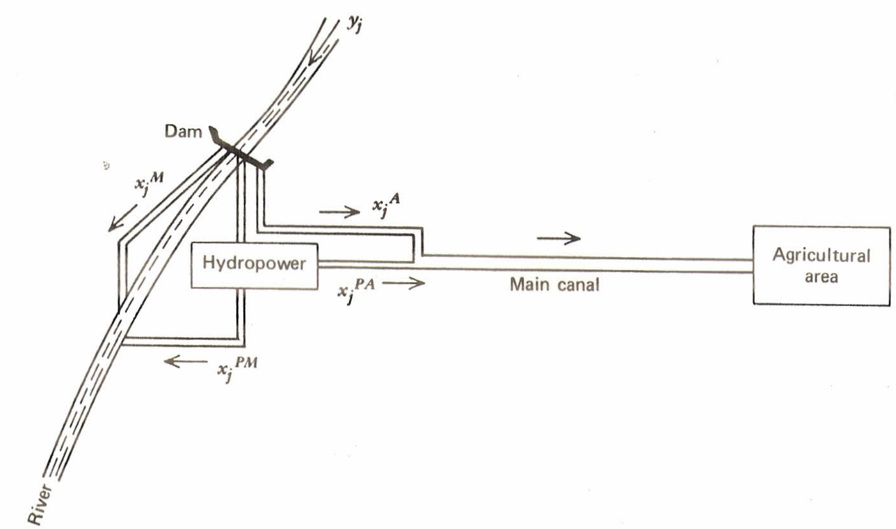


Figure 1.6 A typical river basin.

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 agriculture, power generation, and 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,
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 discounting factor to give 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

$$C_r(S)+\alpha C_l(U) \tag{1.10}$$

where $C_r(S)$ is the cost of the reservoir if its capacity is S, and $C_l(U)$ 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

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

The interested reader may refer to Mobasheri [1968] and Maass et al. [1967] 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, Parikh, and Srinivasan [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.

So 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 at period j, γ is the power conversion factor, and e is the efficiency of the power system. Refer to O'Laoghaire and Himmelblau [1974] for the nature of the function ψ.

Likewise, 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 loss, then 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:

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

where $\mathbf{c}$ and $\mathbf{x}$ are n vectors, $\mathbf{b}$ is an m vector, and $\mathbf{A}=[\mathbf{a}_1,\ldots,\mathbf{a}_n]$ 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, then 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 $c_1, \ldots, c_n$ are not fixed but are random variables. We shall thus assume that $\mathbf{c}$ is a random vector with *mean* $\bar{\mathbf{c}} = (\bar{c}_1, \ldots, \bar{c}_n)'$ and *covariance matrix* $\mathbf{V}$. The objective function, denoted by z, will thus be a random variable with mean $\bar{\mathbf{c}}'\mathbf{x}$ and variance $\mathbf{x}'\mathbf{V}\mathbf{x}$.

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

$$\begin{array}{ll} \text{Maximize} & \bar{\mathbf{c}}'\mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} \le \mathbf{b} \\ & \mathbf{x} \ge \mathbf{0} \end{array}$$

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:

$$\begin{array}{ll} \text{Minimize} & \mathbf{x}'\mathbf{V}\mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} \le \mathbf{b} \\ & \mathbf{x} \ge \mathbf{0} \end{array}$$

which is a quadratic program discussed in Section 11.2.

Satisficing Criteria

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 some research has been done on dealing with such problems (see Zeleny [1974] and Zeleny and Cochrane [1973]). However, there are several other ways of considering both 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 the *aspiration level* or the *satisficing level*. The problem can then be stated as

$$\begin{array}{ll} \text{Minimize} & \mathbf{x}'\mathbf{V}\mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} \le \mathbf{b} \\ & \bar{\mathbf{c}}'\mathbf{x} \ge \bar{z} \\ & \mathbf{x} \ge \mathbf{0} \end{array}$$

which is again a quadratic programming problem.

Another approach that can be adopted is as follows. Let $\alpha = \text{Prob}\,(\mathbf{c}'\mathbf{x} \ge \bar{z})$; that is, α gives the probability that the aspiration level $\bar{z}$ will be attained. Clearly one would like maximize α. Now suppose 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

$$\alpha = \text{Prob}\,(\mathbf{d}'\mathbf{x} + y\mathbf{f}'\mathbf{x} \ge \bar{z})$$
$$= \text{Prob}\left(y \ge \frac{\bar{z} - \mathbf{d}'\mathbf{x}}{\mathbf{f}'\mathbf{x}}\right)$$

if $\mathbf{f}'\mathbf{x} > 0$. Hence, in this case, the problem reduces to

$$\begin{array}{ll} \text{Minimize} & \dfrac{\bar{z} - \mathbf{d}'\mathbf{x}}{\mathbf{f}'\mathbf{x}} \\ \text{subject to} & \mathbf{A}\mathbf{x} \le \mathbf{b} \\ & \mathbf{x} \ge \mathbf{0} \end{array}$$

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

Risk Aversion Model

The above approaches for handling the variance and the expected value of the return do not take into account the risk aversion behavior of individuals. For example, an individual 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. The individual who chooses the expected value of \$100 is more aversive to risk than the individual who might choose the alternative with expected value of \$110. This difference in risk-taking behavior can be taken into account by considering the utility of money for the individual.

For most individuals, 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 individual's utility function and is usually a nondecreasing continuous function. Figure 1.7 gives two typical utility functions for two different individuals. For individual (a), a gain of Δz 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 individual would prefer a lower variance. Such an individual is more aversive to risk than the individual 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-aversive 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

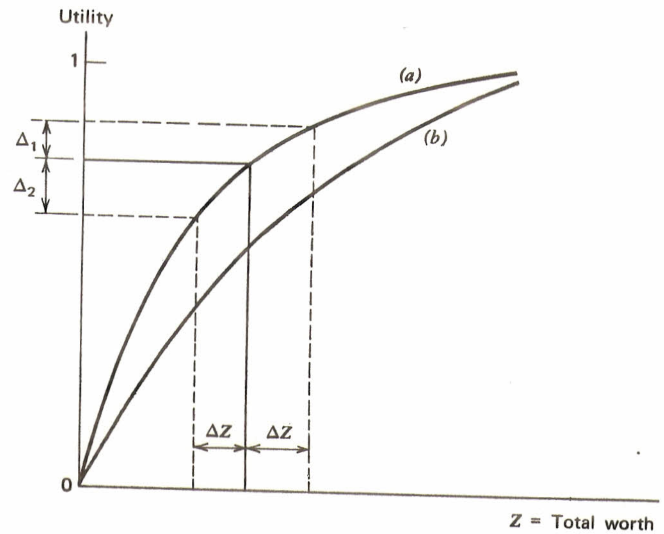


Figure 1.7 Utility functions.

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

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

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

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

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

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

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

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

$$\begin{aligned} \text{Maximize} \quad & k\bar{\mathbf{c}}^t\mathbf{x} - \tfrac{1}{2}k^2\mathbf{x}^t\mathbf{V}\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \le \mathbf{b} \\ & \mathbf{x} \ge \mathbf{0} \end{aligned}$$

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. 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:

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}\,d_{ij} \\ \text{subject to} \quad & \sum_{j=1}^{n} w_{ij} \le c_i \qquad \text{for } i = 1, \ldots, m \\ & \sum_{i=1}^{m} w_{ij} = r_j \qquad \text{for } j = 1, \ldots, n \\ & w_{ij} \ge 0 \qquad \text{for } i = 1, \ldots, m;\ j = 1, \ldots, n \end{aligned}$$

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, such as

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

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

leading to a 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, then the d_{ij}'s are known, and the above problem reduces to a special case of a linear programming problem known as the *transportation problem*.

Exercises

1.1 Consider the following *portfolio selection problem*. An investor must choose a portfolio $\mathbf{x}=(x_1, x_2, \ldots, x_n)'$, where x_j is the proportion of the assets allocated to the jth security. The return on the portfolio has mean $\bar{\mathbf{c}}'\mathbf{x}$ and variance $\mathbf{x}'\mathbf{V}\mathbf{x}$, 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 with 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.

1.2 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

$$0 \le L \le L', \qquad 0 \le W \le W', \qquad 0 \le H \le H'$$

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 the optimal dimensions L, W, and H to minimize the insulation cost.

1.3 Formulate the model for Exercise 1.2 if the storage unit is a cylinder of diameter D and height H.

1.4 An office room of length 60 feet and width 35 feet is to be illuminated by n light bulbs of wattage W_i, $i = 1, \ldots, n$. 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

$$\alpha = 10p, \qquad p = 0, 1, \ldots, 6$$
$$\beta = 5q, \qquad q = 0, 1, \ldots, 7$$

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

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

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 2.6 and 3.2 units is required. The wattage of the bulbs used is between 40 W and 300 W. Assume that the W_i's 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.

c. Select a suitable value of k from the literature. Verify whether the lighting in your classroom is reasonably close to the answer obtained in part b above.

1.5 A household with budget b purchases n commodities. The unit price of commodity j is c_j, and the minimal amount to be purchased of the commodity 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 $l_1, \ldots, l_n$ and $a_1, \ldots, a_n$. 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.

1.6 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 $x_1, \ldots, x_n$ 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 population 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.

1.7 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 μ_j, $j = 1, \ldots, n$. The water quality is examined along the river at m control points. The minimum desired quality improvement at point i is b_i, $i = 1, \ldots, m$. 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 be only 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.

1.8 A 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 into a nonlinear program?

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

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

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

$$3x_1+4x_2\leq 20$$

where x_j is the amount of product j produced. The unit profit for product 1 is normally distributed with mean 3 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?

1.10 Suppose that the demand $d_1,\ldots,d_n$ 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 $j=1,\ldots,n$. If the initial inventory is I_0, formulate the production scheduling problem as a nonlinear program.

1.11 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.

1.12 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 the optimal production quantities Q_1 and Q_2 to minimize the total cost. Write the model for the problem.

b. Now, suppose that shortages are permitted and production need not start when inventory reaches zero level. During the period when inventory is zero and demand is not met, sales are lost. The loss per unit thus incurred is $\$l_j$. On the other hand, if a sale is made, the profit per unit is $\$P_j$. Reformulate the mathematical model.

Notes and References

The advent of high-speed computers has considerably increased our ability to apply iterative procedures in 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, hopefully new theoretical breakthroughs will overcome this problem.

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.

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, Cullum, and Polak [1970], 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.

Mathematical programming has also been successfully used for over a decade 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 on-line every few minutes, with appropriate power adjustments made. For more details, refer to Abou-Taleb et al. [1974], Adams et al. [1972], Beglari and Laughton [1975], Kirchmayer [1958], Sasson [1969a and 1969b], Sasson et al. [1971], and Sasson and Merrill [1974]. The last-mentioned paper gives a summary of many of the applications.

The field of water resources systems analysis has shown a spectacular growth during the last two 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 can be found in Argaman, Shamir, and Spivak [1973], Dajani, Gemmel, and Morlok [1972], Deb and Sarkar [1971], Jacoby [1968], Shamir [1974], 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 1.1, the portfolio allocation problem is briefly introduced.

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 undertaken. 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, Kirby, and Raike [1967], Dantzig [1963], Elmaghraby [1960], Evers [1967], Geoffrion [1967c], Madansky [1962], Mangasarian [1964], Parikh [1970], Sengupta [1972], Sengupta, Portillo-Campbell [1970], Sengupta, Tintner, and Millham [1963], Vajda [1970, 1972], Wets [1966a, 1966b, 1972], Williams [1965, 1966], and Ziemba [1970, 1971, 1974, 1975].