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

WILEY-INTERSCIENCE
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

WILEY-INTERSCIENCE

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

Chapter 1 Introduction 1

1.1 Problem Statement and Basic Definitions 2
1.2 Illustrative Examples 4
1.3 Guidelines for Model Construction 26
Exercises 30
Notes and References 34

Part 1 Convex Analysis 37

Chapter 2 Convex Sets 39

2.1 Convex Hulls 40
2.2 Closure and Interior of a Set 45
2.3 Weierstrass's Theorem 48
2.4 Separation and Support of Sets 50
2.5 Convex Cones and Polarity 62
2.6 Polyhedral Sets, Extreme Points, and Extreme Directions 64
2.7 Linear Programming and the Simplex Method 75
Exercises 86
Notes and References 93

Chapter 3 Convex Functions and Generalizations 97

3.1 Definitions and Basic Properties 98
3.2 Subgradients of Convex Functions 103
3.3 Differentiable Convex Functions 109
3.4 Minima and Maxima of Convex Functions 123
3.5 Generalizations of Convex Functions 134
Exercises 147
Notes and References 159

Part 2 Optimality Conditions and Duality 163

Chapter 4 The Fritz John and Karush–Kuhn–Tucker Optimality Conditions 165

4.1 Unconstrained Problems 166
4.2 Problems Having Inequality Constraints 174
4.3 Problems Having Inequality and Equality Constraints 197
4.4 Second-Order Necessary and Sufficient Optimality Conditions for Constrained Problems 211
Exercises 220
Notes and References 235

Chapter 5 Constraint Qualifications 237

5.1 Cone of Tangents 237
5.2 Other Constraint Qualifications 241
5.3 Problems Having Inequality and Equality Constraints 245
Exercises 250
Notes and References 256

Chapter 6 Lagrangian Duality and Saddle Point Optimality Conditions 257

6.1 Lagrangian Dual Problem 258
6.2 Duality Theorems and Saddle Point Optimality Conditions 263
6.3 Properties of the Dual Function 276
6.4 Formulating and Solving the Dual Problem 286
6.5 Getting the Primal Solution 293
6.6 Linear and Quadratic Programs 298
Exercises 300
Notes and References 313

Part 3 Algorithms and Their Convergence 315

Chapter 7 The Concept of an Algorithm 317

7.1 Algorithms and Algorithmic Maps 317
7.2 Closed Maps and Convergence 319
7.3 Composition of Mappings 324
7.4 Comparison Among Algorithms 329
Exercises 332
Notes and References 340

Chapter 8 Unconstrained Optimization 343

8.1 Line Search Without Using Derivatives 344
8.2 Line Search Using Derivatives 356
8.3 Some Practical Line Search Methods 360
8.4 Closedness of the Line Search Algorithmic Map 363
8.5 Multidimensional Search Without Using Derivatives 365
8.6 Multidimensional Search Using Derivatives 384
8.7 Modification of Newton's Method: Levenberg–Marquardt and Trust Region Methods 398
8.8 Methods Using Conjugate Directions: Quasi-Newton and Conjugate Gradient Methods 402
8.9 Subgradient Optimization Methods 435
Exercises 444
Notes and References 462

Chapter 9 Penalty and Barrier Functions 469

9.1 Concept of Penalty Functions 470
9.2 Exterior Penalty Function Methods 475
9.3 Exact Absolute Value and Augmented Lagrangian Penalty Methods 485
9.4 Barrier Function Methods 501
9.5 Polynomial-Time Interior Point Algorithms for Linear Programming Based on a Barrier Function 509
Exercises 520
Notes and References 533

Chapter 10 Methods of Feasible Directions 537

10.1 Method of Zoutendijk 538
10.2 Convergence Analysis of the Method of Zoutendijk 557
10.3 Successive Linear Programming Approach 568
10.4 Successive Quadratic Programming or Projected Lagrangian Approach 576
10.5 Gradient Projection Method of Rosen 589

10.6 Reduced Gradient Method of Wolfe and Generalized Reduced Gradient Method 602
10.7 Convex–Simplex Method of Zangwill 613
10.8 Effective First- and Second-Order Variants of the Reduced Gradient Method 620
Exercises 625
Notes and References 649

Chapter 11 Linear Complementary Problem, and Quadratic, Separable, Fractional, and Geometric Programming 655

11.1 Linear Complementary Problem 656
11.2 Convex and Nonconvex Quadratic Programming: Global Optimization Approaches 667
11.3 Separable Programming 684
11.4 Linear Fractional Programming 703
11.5 Geometric Programming 712
Exercises 722
Notes and References 745

Appendix A Mathematical Review 751

Appendix B Summary of Convexity, Optimality Conditions, and Duality 765

Bibliography 779

Index 843

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.

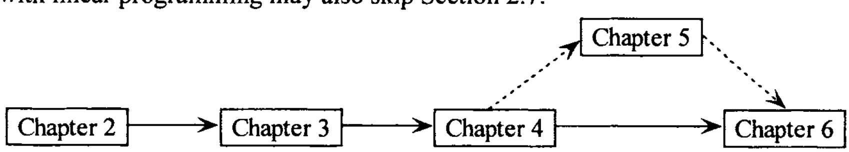
Finally, Chapter 11 deals with some special problems that arise in different applications as well as in the solution of other nonlinear programming problems. In particular, we present the linear complementary, quadratic separable, linear fractional, and geometric programming problems. Methodologies used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

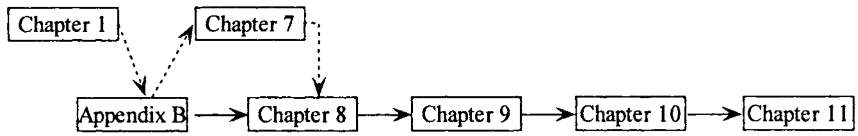
1. *Foundations of Optimization*

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:

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

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

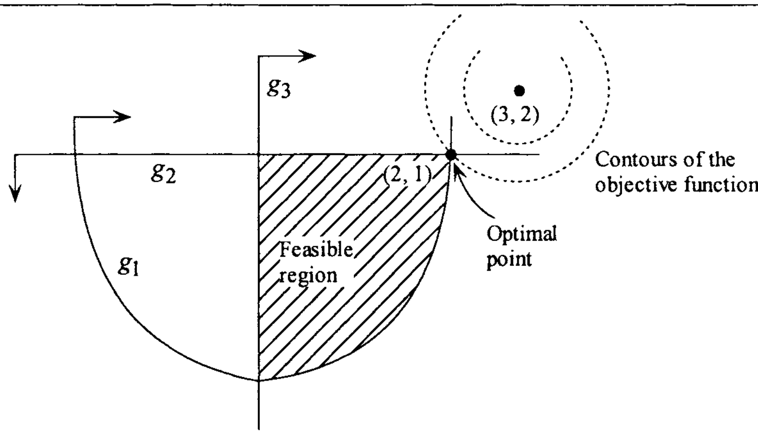


Figure 1.1 Geometric solution of a nonlinear problem.

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 $\boldsymbol{\Psi}$. 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 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:

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

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

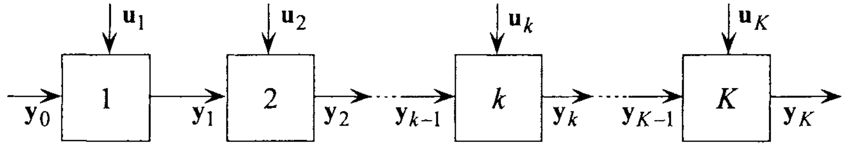


Figure 1.2 Discrete control system.

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:

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

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 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:

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

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:

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

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, 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:

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

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

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

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:

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

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:

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

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 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 $k = 1,\ldots, K$, the above problem can be approximated by the following nonlinear program, where $y_{1,0} = y_{2,0} = 0$:

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

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:

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

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 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:

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

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

B. Structural Design

Structural designers have traditionally 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:

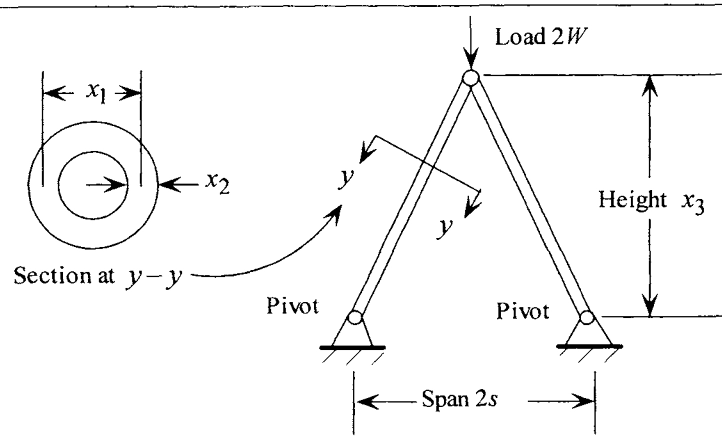


Figure 1.3 Two-bar truss.

$$\begin{aligned}
\text{Minimize} \quad & x_1x_2(s^2+x_3^2)^{1/2} \\
\text{subject to} \quad & x_3 - b_1 \le 0 \\
& x_1 - b_2x_2 \le 0 \\
& W(s^2+x_3^2)^{1/2} - b_3x_1x_2x_3 \le 0 \\
& W(s^2+x_3^2)^{3/2} - b_4x_1x_3(x_1^2+x_2^2) \le 0 \\
& x_1, x_2, x_3 \ge 0.
\end{aligned}$$

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

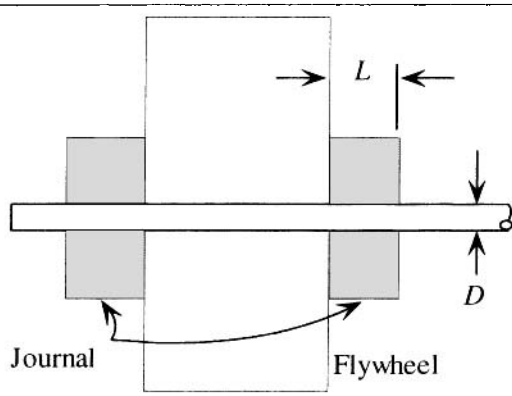


Figure 1.4 Journal bearing assembly.

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 $\hat{h}_0$ 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

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

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: