NONLINEAR PROGRAMMING
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
Dedicated to our parents
# 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
<table>
<thead>
<tr>
<th>Chapter 10</th>
<th>Linear Complementary Problem, and Quadratic, Separable, Fractional, and Geometric Programming</th>
<th>655</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.6</td>
<td>Reduced Gradient Method of Wolfe and Generalized Reduced Gradient Method</td>
<td>602</td>
</tr>
<tr>
<td>10.7</td>
<td>Convex–Simplex Method of Zangwill</td>
<td>613</td>
</tr>
<tr>
<td>10.8</td>
<td>Effective First- and Second-Order Variants of the Reduced Gradient Method</td>
<td>620</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>625</td>
</tr>
<tr>
<td></td>
<td>Notes and References</td>
<td>649</td>
</tr>
<tr>
<td>Chapter 11</td>
<td>Linear Complementary Problem, and Quadratic, Separable, Fractional, and Geometric Programming</td>
<td>655</td>
</tr>
<tr>
<td>11.1</td>
<td>Linear Complementary Problem</td>
<td>656</td>
</tr>
<tr>
<td>11.2</td>
<td>Convex and Nonconvex Quadratic Programming: Global Optimization Approaches</td>
<td>667</td>
</tr>
<tr>
<td>11.3</td>
<td>Separable Programming</td>
<td>684</td>
</tr>
<tr>
<td>11.4</td>
<td>Linear Fractional Programming</td>
<td>703</td>
</tr>
<tr>
<td>11.5</td>
<td>Geometric Programming</td>
<td>712</td>
</tr>
<tr>
<td></td>
<td>Exercises</td>
<td>722</td>
</tr>
<tr>
<td></td>
<td>Notes and References</td>
<td>745</td>
</tr>
<tr>
<td>Appendix A</td>
<td>Mathematical Review</td>
<td>751</td>
</tr>
<tr>
<td>Appendix B</td>
<td>Summary of Convexity, Optimality Conditions, and Duality</td>
<td>765</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td>779</td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td>843</td>
</tr>
</tbody>
</table>
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 \( \ell_1 \) 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 \textit{reformulation–linearization/convexification technique} (RLT) as a \textit{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. \textit{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. \textit{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
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:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m \\
& \quad h_i(x) = 0 \quad \text{for } i = 1, \ldots, \ell \\
& \quad x \in X,
\end{align*}
\]

where \( f, g_1, \ldots, g_m, h_1, \ldots, h_\ell \) are functions defined on \( \mathbb{R}^n \), \( X \) is a subset of \( \mathbb{R}^n \), and \( x \) 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 \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) is called an inequality constraint, and each of the constraints \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \) 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 \( x \in X \) 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 \( \bar{x} \) such that \( f(x) \geq f(\bar{x}) \) for each feasible point \( x \). Such a point \( \bar{x} \) 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 \( g_i(x) \geq 0 \) 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:
Minimize \((x_1 - 3)^2 + (x_2 - 2)^2\)

subject to 
\[ \begin{align*}
  x_1^2 - x_2 - 3 & \leq 0 \\
  x_2 - 1 & \leq 0 \\
  -x_1 & \leq 0.
\end{align*} \]

The objective function and the three inequality constraints are

\[ f(x_1, x_2) = (x_1 - 3)^2 + (x_2 - 2)^2 \]
\[ g_1(x_1, x_2) = x_1^2 - x_2 - 3 \]
\[ g_2(x_1, x_2) = x_2 - 1 \]
\[ g_3(x_1, x_2) = -x_1. \]

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 \textit{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.
Chapter 1

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, \( x^T \) 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 \( \mathbb{R}^n \). 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 \( \theta \). Vector-valued functions are denoted by lowercase Roman or Greek letters, such as \( g \) and \( \Psi \). 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 \( k, \lambda, \) and \( \alpha \).

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 \( y_{k-1} \). A control vector \( u_k \) changes the state of the system from \( y_{k-1} \) to \( y_k \) at the end of period \( k \) according to the following relationship:

\[
y_k = y_{k-1} + \Phi_k (y_{k-1}, u_k) \quad \text{for} \quad k = 1, ..., K.
\]

Given the initial state \( y_0 \), applying the sequence of controls \( u_1, ..., u_K \) would result in a sequence of state vectors \( y_1, ..., y_K \) called the trajectory. This process is illustrated in Figure 1.2.
Figure 1.2 Discrete control system.

A sequence of controls $u_1, \ldots, u_K$ and a sequence of state vectors $y_0, y_1, \ldots, y_K$ are called admissible or feasible if they satisfy the following restrictions:

$$y_k \in Y_k \quad \text{for } k = 1, \ldots, K$$
$$u_k \in U_k \quad \text{for } k = 1, \ldots, K$$
$$\Psi(y_0, \ldots, y_K, u_1, \ldots, u_K) \in D,$$

where $Y_1, \ldots, Y_K$, $U_1, \ldots, U_K$, and $D$ are specified sets, and $\Psi$ 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:

Minimize $\alpha(y_0, y_1, \ldots, y_K, u_1, \ldots, u_K)$
subject to $y_k = y_{k-1} + \phi_k(y_{k-1}, u_k)$ for $k = 1, \ldots, K$
$y_k \in Y_k$ for $k = 1, \ldots, K$
$u_k \in U_k$ for $k = 1, \ldots, K$
$\Psi(y_0, \ldots, y_K, u_1, \ldots, u_K) \in D.$

Combining $y_1, \ldots, y_K, u_1, \ldots, u_K$ as the vector $x$, and by suitable choices of $g$, $h$, 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:

Minimize $\sum_{k=1}^{K} (c_1 u_k^2 + c_2 I_k)$

subject to

$L_k = L_{k-1} + u_k$ for $k = 1, \ldots, K$

$I_k = I_{k-1} + pL_{k-1} - d_k$ for $k = 1, \ldots, K$

$0 \leq L_k \leq b/p$ for $k = 1, \ldots, K$

$I_k \geq 0$ for $k = 1, \ldots, K$,

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, $u$, is to be exerted over the planning horizon $[0, T]$. Given the initial state $y_0$, the relationship between the state vector $y$ and the control vector $u$ is governed by the following differential equation:

$\dot{y}(t) = \Phi[y(t), u(t)]$ for $t \in [0, T]$.  

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

$y(t) \in Y$ for $t \in [0, T]$

$u(t) \in U$ for $t \in [0, T]$

$\Psi(y, u) \in D$.

A typical example of the set $U$ is the collection of piecewise continuous functions on $[0, T]$ such that $a \leq u(t) \leq b$ for $t \in [0, T]$. The optimal control problem can be stated as follows, where the initial state vector $y(0) = y_0$ is given:

Minimize $\int_0^T \alpha[y(t), u(t)] dt$

subject to

$\dot{y}(t) = \Phi[y(t), u(t)]$ for $t \in [0, T]$

$y(t) \in Y$ for $t \in [0, T]$

$u(t) \in U$ for $t \in [0, T]$

$\Psi(y, u) \in D$.  


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 \(\Delta\), such that \(K\Delta = T\). Denoting \(y(k\Delta)\) by \(y_k\) and \(u(k\Delta)\) by \(u_k\), for \(k = 1, \ldots, K\), the above problem can be approximated as follows, where the initial state \(y_0\) is given:

\[
\begin{align*}
\text{Minimize} & \quad \sum_{k=1}^{K} \alpha(y_k, u_k) \\
\text{subject to} & \quad y_k = y_{k-1} + \Delta\phi(y_{k-1}, u_k) \quad \text{for } k = 1, \ldots, K \\
& \quad y_k \in Y \quad \text{for } k = 1, \ldots, K \\
& \quad u_k \in U \quad \text{for } k = 1, \ldots, K \\
& \quad \Psi(y_0, \ldots, y_K, u_1, \ldots, u_K) \in D.
\end{align*}
\]

**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

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

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:

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

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:

\[
\begin{align*}
\text{Minimize} & \quad \int_0^T |u(t)| \dot{y}_2(t) \, dt \\
\text{subject to} & \quad \dot{y}_1(t) = y_2(t) \quad \text{for } t \in [0, T] \\
& \quad m\dot{y}_2(t) = u(t) - mg \quad \text{for } t \in [0, T] \\
& \quad |u(t)| \leq b \quad \text{for } t \in [0, T] \\
& \quad y_1(T) = \bar{y},
\end{align*}
\]
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 \( \ell \). 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 \( y_{1,0} = y_{2,0} = 0\):

\[
\text{Minimize } \sum_{k=1}^{K} |u_k| y_{2,k}
\]

subject to

\[
\begin{align*}
y_{1,k} - y_{1,k-1} &= y_{2,k-1} & \text{for } k = 1, \ldots, K \\
m(y_{2,k} - y_{2,k-1}) &= u_k - mg & \text{for } k = 1, \ldots, K \\
|u_k| &\leq b & \text{for } k = 1, \ldots, K \\
y_{1,K} &= \bar{y}
\end{align*}
\]

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, \( |\dot{y}(t)| \leq b_1 \). 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, \( |\ddot{y}(t)| \leq b_2 \). Furthermore, the end conditions \( y(0) = a \) and \( y(T) = b \) must be observed. The problem can thus be stated as follows:

\[
\text{Minimize } \int_0^T |y(t) - c(t)| \, dt
\]

subject to

\[
\begin{align*}
|\dot{y}(t)| &\leq b_1 & \text{for } t \in [0, T] \\
|\ddot{y}(t)| &\leq b_2 & \text{for } t \in [0, T] \\
y(0) &= a \\
y(T) &= b
\end{align*}
\]

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 \( \ell \). 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:
Introduction

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

\[
\begin{align*}
\text{Minimize} & \quad \sum_{k=1}^{K} |y_{1,k} - c_k| \\
\text{subject to} & \quad y_{1,k} - y_{1,k-1} = y_{2,k-1} \quad \text{for } k = 1, ..., K \\
& \quad -b_1 \leq y_{2,k} \leq b_1 \quad \text{for } k = 0, ..., K - 1 \\
& \quad -b_2 \leq y_{2,k} - y_{2,k-1} \leq b_2 \quad \text{for } k = 1, ..., K - 1 \\
& \quad y_{1,0} = a \\
& \quad y_{1,K} = b.
\end{align*}
\]
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 $\rho$ 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, $x_3 \leq b_1$.

2. The ratio of the diameter of the tube to the thickness of the tube must not exceed $b_2$; that is, $x_1/x_2 \leq b_2$.

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:

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

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:

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

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

![Figure 1.3 Two-bar truss.](image)
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

\[
\begin{align*}
\text{Minimize} & \quad x_1x_2(s^2 + x_3^2)^{1/2} \\
\text{subject to} & \quad x_3 - b_1 \leq 0 \\
& \quad x_1 - b_2x_2 \leq 0 \\
& \quad W(s^2 + x_3^2)^{1/2} - b_3x_1x_2x_3 \leq 0 \\
& \quad W(s^2 + x_3^2)^{3/2} - b_4x_1x_3(x_1^2 + x_2^2) \leq 0 \\
& \quad x_1, x_2, x_3 \geq 0.
\end{align*}
\]

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

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

where $h_0$ is the minimum oil film thickness to prevent metal-to-metal contact and $\delta$ 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:

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

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

$$\theta = \frac{1}{k_1D}.$$

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 $\omega$ 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

$$\phi(e) = \frac{e}{(1-e^2)^{3/2}} \left[ \pi^2 (1-e^2) + 16e^2 \right]^{1/2}.$$ 

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

Thus, if $\delta$, $\hat{h}_0$, 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 $\alpha$. The model is thus given by:
Introduction

\[
\begin{align*}
\text{Minimize} & \quad \frac{\omega}{\delta \sqrt{1-e^2}} D^3 L \\
\text{subject to} & \quad \frac{1}{k_1 D} \leq \alpha \\
& \quad 2 \frac{k_2 \omega}{\delta^2} DL^3 \phi \left(1 - \frac{h_0}{\delta}\right) \geq W \\
& \quad \hat{h}_0 \leq h_0 \leq \delta \\
& \quad 0 \leq 1 - \frac{h_0}{\delta} \leq \hat{e} \\
& \quad D \geq 0 \\
& \quad L \geq 0.
\end{align*}
\]

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

D. Electrical Networks

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

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

1. A voltage source that maintains a constant branch voltage \(v_s\) irrespective of the branch current \(c_s\). Such a device absorbs power equal to \(-v_sc_s\).

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

\[
c_d \geq 0, \quad v_d \geq 0, \quad v_d c_d = 0. \tag{1.1}
\]

3. A resistor that consumes power and whose branch current \(c_r\) and branch voltage \(v_r\) are related by...
where \( r \) is the resistance of the resistor. The power consumed is given by

\[
-v_r c_r = \frac{v_r^2}{r} = r c_r^2.
\]  

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

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

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

For a network having several voltage sources, diodes, and resistors, let \( N_S \) denote the node-branch incidence matrix for all the branches having voltage sources, \( N_D \) denote the node-branch incidence matrix for all branches having diodes, and \( N_R \) denote the node-branch incidence matrix for all branches having resistors. Then, without loss of generality, we can partition \( N \) as

\[
N = [N_S, N_D, N_R].
\]

Similarly, the column vector \( c \), representing the branch currents, can be partitioned as

\[
c' = [c'_S, c'_D, c'_R].
\]

![Figure 1.5 Typical electrical devices in a circuit.](image)
and the column vector \( v \), representing the branch voltages, can be written as

\[
v' = [v'_S, v'_D, v'_R].
\]

Associated with each node \( i \) is a node potential \( p_i \). The column vector \( p \), representing node potentials, can be written as

\[
p' = [p'_S, p'_D, p'_R].
\]

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 \( Nc = 0 \), or

\[
N'_S c_S + N'_D c_D + N'_R c_R = 0.
\]  
(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 \( N'c = v \), or

\[
N'_S p = v_S \\
N'_D p = v_D \\
N'_R p = v_R.
\]  
(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

\[
v_D \geq 0, \quad c_D \geq 0, \quad v'_D c_D = 0,
\]  
(1.6)

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

\[
v_R = -Rc_R,
\]  
(1.7)

where \( R \) is a diagonal matrix whose diagonal elements are the resistance values.

Thus, (1.4) – (1.7) represent the equilibrium conditions of the circuit, and we wish to find \( v_D, v_R, c, \) and \( p \) satisfying these conditions.

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

Minimize \[
\frac{1}{2} c' R c_R - v'_S c_S
\]
subject to \[
N'_S c_S + N'_D c_D + N'_R c_R = 0
\]
\[-c_D \leq 0.
\]
Here we wish to determine the branch currents \( c_S, c_D, \) and \( c_R \) 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

\[
\begin{align*}
N'_{S}u - v_S &= 0 \\
N'_{D}u - Iu_0 &= 0 \\
N'_{R}u + Rc_R &= 0 \\
N_{S}c_S + N_{D}c_D + N_{R}c_R &= 0 \\
c_D'u_0 &= 0 \\
c_D, u_0 &\geq 0,
\end{align*}
\]

where \( u \) and \( u_0 \) are column vectors representing the Lagrangian multipliers. It can readily be verified that letting \( v_D = u_0 \), \( p = u \), and noting (1.7), the conditions above are precisely the equilibrium conditions (1.4) – (1.7). Note that the Lagrangian multiplier vector \( u \) is precisely the node potential vector \( p \).

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

Maximize \(-\frac{1}{2}v'^T_RGv_R\)

subject to \( N'_S p = v_S \)

\( N'_D p - v_D = 0 \)

\( N'_R p - v_R = 0 \)

\( v_D \geq 0 \).

Here, \( v'^T_RGv_R \) is the power absorbed by the resistors, and we wish to find the branch voltages \( v_D \) and \( v_R \) and the potential vector \( p \).

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

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

\[
\frac{1}{2}c'_RRe_R + \frac{1}{2}v'^T_RGv_R - v'_S c_S = 0.
\]

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

\[
v'_Rc_R + v'_Dc_D + v'_Sc_S = 0,
\]

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

E. Water Resources Management

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

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

There are two classes of variables associated with the problem:

1. **Design variables**: What should be the optimal capacity $S$ of the reservoir, the capacity $U$ of the canal supplying agricultural water, and the capacity $E$ of the power plant?

2. **Operational variables**: How much water should be released for agricultural power generation and for other purposes?

From Figure 1.6, the following operational variables can readily be identified for the $j$th period:

- $x_j^A = \text{water released from the dam for agriculture}$
- $x_j^{PA} = \text{water released for power generation and then for agricultural use}$

---

![Figure 1.6 Typical river basin.](image)
\( x_j^{PM} \) = 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

\[
C(E) + \sum_{j=1}^{N} \beta_j \hat{C}_e(E),
\]

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, \( \beta_j \) is a discount factor that gives the present worth of the cost in period \( j \). See Mobasheri [1968] for the nature of the functions \( C(E) \) and \( \hat{C}_e(E) \).

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

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

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

**Reservoir and Canal:** The discounted capital costs are given by

\[
C_r(S) + \alpha C_\ell(U),
\]

where \( C_r(S) \) is the cost of the reservoir if its capacity is \( S \), and \( C_\ell(U) \) is the capital cost of the main canal if its capacity is \( U \). Here \( \alpha \) 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

\[
\sum_{j=1}^{N} \beta_j \left[ \hat{C}_r(S) + \hat{C}_\ell(U) \right].
\]
The interested reader may refer to Maass et al. [1967] and Mobasheri [1968] for a discussion on the nature of the functions discussed here.

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

$$\sum_{j=1}^{N} \beta_j R(x_j^A + x_j^P).$$  \hspace{1cm} (1.12)

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

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

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

$$f_j \leq (x_j^{PM} + x_j^{PA}) \Psi(s_j) \gamma e,$$  \hspace{1cm} (1.13)

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

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

$$f_j \leq \alpha_j E e H_j,$$  \hspace{1cm} (1.14)

where $\alpha_j$ 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,

$$E' \leq E \leq E''.$$  \hspace{1cm} (1.15)

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

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

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

$$S \geq s_j.$$  \hspace{1cm} (1.17)
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

$$x_j^M + x_j^{PM} \geq M_j.$$  
(1.19)

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

$$x_j^A + x_j^{PA} \leq U.$$  
(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:

Maximize $c'x$

subject to $Ax \leq b$

$$x \geq 0,$$

where $c$ and $x$ are $n$-vectors, $b$ is an $m$-vector, and $A = [a_1, ..., a_n]$ 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 $b$. Column $a_j$ of $A$ 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 $a_j x_j$ of the available resources; hence the constraint, $Ax = \sum_{j=1}^{n} a_j x_j \leq b$. If the unit profit of activity $j$ is $c_j$, the total profit is $\sum_{j=1}^{n} c_j x_j = c'x$. Thus, the problem can be interpreted as finding the best way of allocating the resource vector $b$ to the various available activities so that the total profit is maximized.

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

If we want to maximize the expected value of $z$, we must solve the following problem:
Maximize $\bar{c}'x$

subject to $Ax \leq b$

$$x \geq 0,$$

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

Minimize $x'Vx$

subject to $Ax \leq b$

$$x \geq 0,$$

which is a quadratic program as discussed in Section 11.2.

**Satisficing Criteria and Chance Constraints**

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

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

Minimize $x'Vx$

subject to $Ax \leq b$

$$\bar{c}'x \geq \bar{z}$$

$$x \geq 0,$$

which is again a quadratic programming problem.

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

$$\alpha = \text{Prob}(d'x + yf'x \geq \bar{z})$$

$$= \text{Prob}\left\{ y \geq \frac{\bar{z} - d'x}{f'x} \right\}$$

if $f'x > 0$. Hence, in this case, the problem of maximizing $\alpha$ reduces to:
Minimize \( \frac{\bar{z} - d'x}{f'x} \)
subject to \( Ax \leq b \)
\( x \geq 0. \)

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

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

\[
\text{Prob}(c'x \geq \bar{z}) = \text{Prob}\left(y \geq \frac{\bar{z} - d'x}{f'x}\right) \geq q.
\]

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

\[
\frac{\bar{z} - d'x}{f'x} \leq \phi_q \quad \text{or} \quad d'x + \phi_q f'x \geq \bar{z}.
\]

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

**Risk Aversion Model**

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

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

Our new utility of $\Delta_1$, and a loss of $\Delta_2$ decreases the utility by $\Delta_2$. Since $\Delta_2$ is larger than $\Delta_1$, this person would prefer a lower variance. Such a person is more averse to risk than a person whose utility function is as in (b) in Figure 1.7.

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

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

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

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

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

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

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

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

$$= 1 - \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left( -\frac{1}{2} \left( \frac{z - \bar{z} + k\sigma^2}{\sigma} \right)^2 \right) \exp\left( -k\bar{z} + \frac{1}{2} k^2 \sigma^2 \right) \, dz.$$
\begin{equation}
\begin{aligned}
&= 1 - \exp\left(-k\bar{z} + \frac{1}{2}k^2 \sigma^2 \right) \\
&= 1 - \exp\left(-k\bar{z} + \frac{1}{2}k^2 \sigma^2 \right).
\end{aligned}
\end{equation}

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

Maximize \( k\bar{c}'x - \frac{1}{2}k^2 x'\mathbf{V}x \)

subject to \( Ax \leq b \)

\( x \geq 0 \).

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

**G. Location of Facilities**

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

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

\[
(x_i, y_i) \quad \text{unknown location of warehouse } i \text{ for } i = 1, \ldots, m
\]

\[
c_i \quad \text{capacity of warehouse } i \text{ for } i = 1, \ldots, m
\]

\[
(a_j, b_j) \quad \text{known location of market } j \text{ for } j = 1, \ldots, n
\]

\[
r_j \quad \text{known demand at market } j \text{ for } j = 1, \ldots, n
\]

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

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

The problem of locating the warehouses and determining the shipping pattern can be stated as follows:
Minimize \[ \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} d_{ij} \]

subject to \[ \sum_{j=1}^{n} w_{ij} \leq c_i \quad \text{for } i = 1, \ldots, m \]
\[ \sum_{i=1}^{m} w_{ij} = r_j \quad \text{for } j = 1, \ldots, n \]
\[ w_{ij} \geq 0 \quad \text{for } i = 1, \ldots, m; j = 1, \ldots, n. \]

Note that both \( w_{ij} \) and \( d_{ij} \) are to be determined, and hence, the above problem is a nonlinear programming problem. Different measures of distance can be chosen, using the rectilinear, Euclidean, or \( \ell_p \) norm metrics, where the value of \( p \) could be chosen to approximate particular city travel distances. These are given respectively by
\[
\begin{align*}
    d_{ij} &= |x_i - a_j| + |y_i - b_j| \\
    d_{ij} &= \left[ (x_i - a_j)^2 + (y_i - b_j)^2 \right]^{1/2} \\
    d_{ij} &= \left[ (x_i - a_j)^p + (y_i - b_j)^p \right]^{1/p}.
\end{align*}
\]

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

H. Miscellaneous Applications

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

The modeling process is concerned with the construction of a mathematical abstraction of a given problem that can be analyzed to produce meaningful answers that guide the decisions to be implemented. Central to this process is the identification or the formulation of the problem. By the nature of human activities, a problem is seldom isolated and crisply defined, but rather, interacts with various other problems at the fringes and encompasses various details obfuscated by uncertainty. For example, a problem of scheduling jobs on machines interacts with the problems of acquiring raw materials, forecasting uncertain demand, and planning for inventory storage and dissipation; and it must contend with machine reliability, worker performance and absenteeism, and insertions of spurious or rush-jobs. A modeler must therefore identify the particular scope and aspect of the problem to be explicitly considered in formulating the problem, and must make suitable simplifying assumptions so that the resulting model is a balanced compromise between representability and mathematical tractability.

The model, being only an abstraction of the real problem, will yield answers that are only as meaningful as the degree of accuracy with which it represents the actual physical system. On the other hand, an unduly complicated model might be too complex to be analyzed mathematically for obtaining any credible solution for consideration at all! This compromise, of course, need not be achieved at a single attempt. Often, it is instructive to begin with a simpler model representation, to test it to gain insights into the problem, and then to guide the direction in which the model should be further refined to make it more representative while maintaining adequate tractability. While accomplishing this, it should be borne in mind that the answers from the model are meant to provide guidelines for making decisions rather than to replace the decision maker. The model is only an abstraction of reality and is not necessarily an equivalent representation of reality itself. At the same time, these guidelines need to be well founded and meaningful. Moreover, one important function of a model is to provide more information on system behavior through sensitivity analyses, in which the response of the system is studied under various scenarios related to perturbations in different problem parameters. To obtain reliable insights through such an analysis, it is important that a careful balance be struck between problem representation and tractability.

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

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

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

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

or as

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

or

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

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

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

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

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

In the modeling process it is also useful to distinguish between hard constraints, which must necessarily be satisfied without any compromise, and soft constraints, for which mild violations can be tolerated, albeit at some incurred cost. For example, the expenditure $g(x)$ for some activity vector $x$ might be required to be no more than a budgeted amount $B$, but violations within limits might be permissible if economically justifiable. Hence, this constraint can be modeled as $g(x) - B = y^+ - y^-$, where $y^+$ and $y^-$ are nonnegative variables, and where the "violation" $y^+$ is bounded above by a limit on the capital that can be borrowed or raised and, accordingly, also accompanied by a cost term $c(y^+)$ in the objective function. Such constraints are also referred to as elastic constraints because of the flexibility they provide.
It is insightful to note that permitting mild violations in some constraints, if tolerable, can have a significant impact on the solution obtained. For example, imposing a pair of constraints \( h_1(x) = 0 \) and \( h_2(x) = 0 \) as hard constraints might cause the feasible region defined by their intersection to be far removed from attractively valued solutions, while such solutions only mildly violate these constraints. Hence, by treating them as soft constraints and rewriting them as \(-\Delta_i \leq h_i(x) \leq \Delta_i\), where \( \Delta_i \) is a small positive tolerance factor for \( i = 1, 2 \), we might be able to obtain far better solutions which, from a managerial viewpoint, compromise more judiciously between solution quality and feasibility. These concepts are related to goal programming (see Ignizio [1976]), where the soft constraints represent goals to be attained, along with accompanying penalties or rewards for under- or over-achievements.

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

[1.1] Consider the following nonlinear programming problem:

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

a. Sketch the feasible region and the contours of the objective function. Hence, identify the optimum graphically on your sketch.

b. Repeat part a by replacing minimization with maximization in the problem statement.

[1.2] Suppose that the daily demand for product \( j \) is \( d_j \) for \( j = 1, 2 \). The demand should be met from inventory, and the latter is replenished from production whenever the inventory reaches zero. Here, the production time is assumed to be insignificant. During each product run, \( Q_j \) units can be produced at a fixed setup cost of \( k_j \) and a variable cost of \( c_j Q_j \). 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 \( TD_j k_j Q_j + Tc_j d_j + TQ_j h_j/2 \). Adequate storage area for handling the maximum inventory \( Q_j \) has to be reserved for each product \( j \). Each unit of product \( j \) needs \( s_j \) square feet of storage space, and the total space available is \( S \).

a. We wish to find optimal production quantities \( Q_1 \) and \( Q_2 \) to minimize the total cost. Construct a model for this problem.

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

[1.3] A manufacturing firm produces four different products. One of the necessary raw materials is in short supply, and only \( R \) pounds are available. The selling price of product \( i \) is \( S_i \) per pound. Furthermore, each pound of product \( i \) uses \( a_i \) pounds of the critical raw material. The variable cost, excluding the raw material cost, of producing \( x_i \) pounds of product \( i \) is \( k_i x_i^2 \), where \( k_i > 0 \) is known. Develop a mathematical model for the problem.
[1.4] 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.

[1.5] An office room of length 70 feet and width 45 feet is to be illuminated by $n$ light bulbs of wattage $W_i$, $i = 1, \ldots, n$. 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 $i$th bulb. To ensure adequate lighting, illumination is checked at the working surface level at grid points of the form $(a, p)$, where

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

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

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

- 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.
- Construct a mathematical model similar to that of part a, with the added restriction that all bulbs must be of the same wattage.

[1.6] Consider the following portfolio selection problem. An investor must choose a portfolio $x = (x_1, x_2, \ldots, x_n)^T$, where $x_j$ is the proportion of the assets allocated to the $j$th security. The return on the portfolio has mean $\bar{e}^T x$ and variance $x^T V x$, where $\bar{e}$ is the vector denoting mean returns and $V$ is the matrix of covariances of the returns. The investor would like to increase his or her expected return while decreasing the variance and hence the risk. A portfolio is called efficient if there exists no other portfolio having a larger expected return and a smaller variance. Formulate the problem of finding an efficient portfolio, and suggest some procedures for choosing among efficient portfolios.
A household with budget $b$ purchases $n$ commodities. The unit price of commodity $j$ is $c_j$, and the minimal amount of the commodity to be purchased is $\ell_j$. After the minimal amounts of the $n$ products are consumed, a function $a_j$ of the remaining budget is allocated to commodity $j$. The behavior of the household is observed over $m$ months for the purpose of estimating $\ell_1, \ldots, \ell_n$, 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.

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

$$0 \leq L \leq L', \quad 0 \leq W \leq W', \quad \text{and} \quad 0 \leq H \leq H'.$$

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

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

Suppose that the demand for a certain product is a normally distributed random variable with mean 150 and variance 49, and that the production function is given by $p(x) = a^T x$, where $x$ represents a set of $n$ activity levels. Formulate the chance constraint that the probability of production falling short of demand by more than 5 units should be no more than 1% as a linear constraint.
Consider a linear program to minimize $c'x$ subject to $Ax \leq b$, $x \geq 0$. Suppose that the components $c_j$ of the vector $c$ are random variables distributed independently of each other and of the $x$-variables, 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 $\bar{c}'x$ subject to $Ax \leq b$, $x \geq 0$, where $\bar{c} = (\bar{c}_1, \ldots, \bar{c}_n)'$.

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

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

where $x_j$ is the amount of product $j$ produced. The unit profit for product 1 is normally distributed with mean 4 and variance 2. The unit profit for product 2 is given by a $\chi^2$-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?

Consider the following problem of a regional effluent control along a river. Currently, $n$ manufacturing facilities dump their refuse into the river. The current rate of dumping by facility $j$ is $\mu_j$, $j = 1, \ldots, n$. 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 only be required to remove small amounts of waste or none at all. Reformulate the problem so that a measure of equity among the sources is attained.

A steel company manufactures crankshafts. Previous research indicates that the mean shaft diameter may assume the value $p_1$ or $p_2$, where $p_2 > p_1$. Furthermore, the probability that the mean is equal to $p_1$ is $p$. To test whether the mean is $p_1$ or $p_2$, a sample of size $n$ is chosen, and the diameters $x_1, \ldots, x_n$ are recorded. If $\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$ is less than or equal to $K$, the hypothesis $\mu = p_1$ is accepted; otherwise, the hypothesis $\mu = p_2$ is accepted. Let $f(\bar{x} | p_1)$ and $f(\bar{x} | p_2)$ be the probability density functions of the sample mean if the popula-
tion mean is $\mu_1$ and $\mu_2$, respectively. Furthermore, suppose that the penalty cost of accepting $\mu = \mu_1$ when $\mu = \mu_2$ is $\alpha$ and that the penalty cost of accepting $\mu = \mu_2$ when $\mu = \mu_1$ is $\beta$. Formulate the problem of choosing $K$ such that the expected total cost is minimized. Show how the problem could be reformulated as a nonlinear program.

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

Notes and References

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

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

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

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

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

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

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

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

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

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

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

Finally, we mention that various modeling languages, such as GAMS (see Brooke et al., 1985), LINGO (see Cunningham and Schrage, 1989), and AMPL (see Fourer et al., 1990), are available to assist in the implementation of models and algorithms. Various nonlinear programming software packages, such as MINOS (see Murtagh and Saunders, 1982), GINO (see Liebman et al., 1986), GRG2 (see Lasdon et al., 1978), CONOPT (see Drud, 1985), SQP (see Mahidhara and Lasdon, 1990), LSGRG (see Smith and Lasdon, 1992), BARON (see Sahinidis, 1996), and LGO (see Pintér, 2000, 2001), among others, are also available to facilitate implementation. (The latter two are global optimizer software packages—see Chapter 11.) For a general discussion on algorithms and software evaluation for nonlinear optimization, see DiPillo and Murli [2003].
Part 1

Convex Analysis
Chapter 2 Convex Sets

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

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

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

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

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

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

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

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

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

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

2.1.1 Definition

A set $S$ in $\mathbb{R}^n$ is said to be convex if the line segment joining any two points of the set also belongs to the set. In other words, if $x_1$ and $x_2$ are in $S$, then $\lambda x_1 + (1 - \lambda)x_2$, must also belong to $S$ for each $\lambda \in [0,1]$. Weighted averages of the form $\lambda x_1 + (1 - \lambda)x_2$, where $\lambda \in [0,1]$, are referred to as convex combinations of $x_1$ and $x_2$. Inductively, weighted averages of the form $\sum_{j=1}^{k} \lambda_j x_j$, where $\sum_{j=1}^{k} \lambda_j = 1$, $\lambda_j \geq 0$, $j = 1, \ldots, k$, are also called convex combinations of $x_1, \ldots, x_k$. In this definition, if the nonnegativity conditions on the multipliers $\lambda_j$ is dropped, $j = 1, \ldots, k$, the combination is known as an affine combination. Finally, a combination $\sum_{j=1}^{k} \lambda_j x_j$ where the multipliers $\lambda_j$, $j = 1, \ldots, k$, are simply required to be in $\mathbb{R}$, is known as a linear combination.

Figure 2.1 illustrates the notion of a convex set. Note that in Figure 2.1b, the line segment joining $x_1$ and $x_2$ does not lie entirely in the set.

The following are examples of convex sets:

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

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

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

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

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

   This set is the intersection of two half-spaces. In general, the set $S = \{x : Ax \leq b\}$ is a convex set, where $A$ is an $m \times n$ matrix and $b$ is an $m$-vector. This set is the intersection of $m$ half-spaces and is usually called a polyhedral set.
Figure 2.1 Convex and nonconvex sets.

4. \[ S = \{(x_1, x_2) : x_2 \geq |x_1| \} \subseteq \mathbb{R}^2. \]

This set represents a convex cone in \( \mathbb{R}^2 \) and is treated more fully in Section 2.4.

5. \[ S = \{(x_1, x_2) : x_1^2 + x_2^2 \leq 4 \} \subseteq \mathbb{R}^2. \]

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

6. \[ S = \{x : x \text{ solves Problem P below}\}: \]

Problem P: Minimize \( c'x \)

subject to \( Ax = b \)

\( x \geq 0. \)

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

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

**2.1.2 Lemma**

Let \( S_1 \) and \( S_2 \) be convex sets in \( \mathbb{R}^n \). Then:

1. \( S_1 \cap S_2 \) is convex.
2. \( S_1 \oplus S_2 = \{x_1 + x_2 : x_1 \in S_1, x_2 \in S_2\} \) is convex.
3. \( S_1 \ominus S_2 = \{x_1 - x_2 : x_1 \in S_1, x_2 \in S_2\} \) is convex.
Chapter 2

Convex Hulls

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

2.1.3 Definition

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

\[
x = \sum_{j=1}^{k} \lambda_j x_j
\]

\[
\sum_{j=1}^{k} \lambda_j = 1
\]

\[
\lambda_j \geq 0 \quad \text{for } j = 1, \ldots, k,
\]

where \( k \) is a positive integer and \( x_1, \ldots, x_k \in S \).

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

2.1.4 Lemma

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

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

Figure 2.2 Convex hulls.
We have discussed above the convex hull of an arbitrary set $S$. The convex hull of a finite number of points leads to the definitions of a polytope and a simplex.

2.1.5 Definition

The convex hull of a finite number of points $x_1, \ldots, x_{k+1}$ in $R^n$ is called a polytope. If $x_1, x_2, \ldots, x_k$, and $x_{k+1}$ are affinely independent, which means that $x_2 - x_1, x_3 - x_1, \ldots, x_{k+1} - x_1$ are linearly independent, then $\text{conv}(x_1, \ldots, x_{k+1})$, the convex hull of $x_1, \ldots, x_{k+1}$, is called a simplex having vertices $x_1, \ldots, x_{k+1}$.

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

Carathéodory Theorem

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

2.1.6 Theorem

Let $S$ be an arbitrary set in $R^n$. If $x \in \text{conv}(S)$, $x \in \text{conv}(x_1, \ldots, x_{n+1})$, where $x_j \in S$ for $j = 1, \ldots, n + 1$. In other words, $x$ can be represented as

![Figure 2.3 Polytope and simplex.](image-url)
Chapter 2

\[ x = \sum_{j=1}^{n+1} \lambda_j x_j \]

\[ \sum_{j=1}^{n+1} \lambda_j = 1 \]

\[ \lambda_j \geq 0 \quad \text{for } j = 1, \ldots, n+1 \]

\[ x_j \in S \quad \text{for } j = 1, \ldots, n+1. \]

**Proof**

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

Toward this end, note that \( x_2 - x_1, x_3 - x_1, \ldots, x_k - x_1 \) are linearly dependent. Thus, there exist scalars \( \mu_2, \mu_3, \ldots, \mu_k \) not all zero such that \( \sum_{j=2}^{k} \mu_j (x_j - x_1) = 0 \). Letting \( \mu_1 = -\sum_{j=2}^{k} \mu_j \), it follows that \( \sum_{j=1}^{k} \mu_j x_j = 0 \), \( \sum_{j=1}^{k} \mu_j = 0 \), and not all the \( \mu_j \) values are equal to zero. Note that at least one \( \mu_j \) is larger than zero. Then

\[ x = \sum_{j=1}^{k} \lambda_j x_j + 0 = \sum_{j=1}^{k} \lambda_j x_j - \alpha \sum_{j=1}^{k} \mu_j x_j = \sum_{j=1}^{k} (\lambda_j - \alpha \mu_j) x_j \]

for any real \( \alpha \). Now, choose \( \alpha \) as follows:

\[ \alpha = \min_{1 \leq j \leq k} \left\{ \frac{\lambda_j}{\mu_j} : \mu_j > 0 \right\} = \frac{\lambda_i}{\mu_i} \quad \text{for some } i \in \{1, \ldots, k\}. \]

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

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

2.2.1 Definition

Let \( S \) be an arbitrary set in \( \mathbb{R}^n \). A point \( x \) is said to be in the **closure** of \( S \), denoted by \( \text{cl} \ S \), if \( S \cap N_\varepsilon(x) \neq \emptyset \) for every \( \varepsilon > 0 \). If \( S = \text{cl} \ S \), \( S \) is called **closed**.

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

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

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

There is another equivalent definition of a closed set, which is often important from the viewpoint of demonstrating that a set is closed. This definition is based on sequences of points contained in \( S \) (review Appendix A for related mathematical concepts). A set \( S \) is closed if and only if for any convergent sequence of points \( \{ x_k \} \) contained in \( S \) with limit point \( \bar{x} \), we also have that \( \bar{x} \in S \). The equivalence of this and the previous definition of
closedness is easily seen by noting that the limit point \( \bar{x} \) of any convergent sequence of points in \( S \) must either lie in the interior or on the boundary of \( S \), since otherwise, there would exist an \( \varepsilon > 0 \) such that \( \{ x : \| x - \bar{x} \| < \varepsilon \} \cap S = \emptyset \), contradicting that \( \bar{x} \) is the limit point of a sequence contained in \( S \). Hence, if \( S \) is closed, \( \bar{x} \in S \). Conversely, if \( S \) satisfies the sequence property above, it is closed, since otherwise there would exist some boundary point \( \bar{x} \) not contained in \( S \). But by the definition of a boundary point, the set \( N_{\varepsilon_k}(\bar{x}) \cap S \neq \emptyset \) for each \( k = 1, 2, \ldots \), where \( 0 < \varepsilon < 1 \) is some scalar. Hence, selecting some \( x_k \in N_{\varepsilon_k}(\bar{x}) \cap S \) for each \( k = 1, 2, \ldots \), we will have \( \{ x_k \} \subseteq S \); and clearly \( \{ x_k \} \to \bar{x} \), which means that we must have \( \bar{x} \in S \) by our hypothesis. This is a contradiction.

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

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

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

**2.2.2 Theorem**

Let \( S \) be a convex set in \( \mathbb{R}^n \) with a nonempty interior. Let \( x_1 \in \text{cl} \ S \) and \( x_2 \in \text{int} \ S \). Then \( \lambda x_1 + (1 - \lambda)x_2 \in \text{int} \ S \) for each \( \lambda \in (0, 1) \).

**Proof**

Since \( x_2 \in \text{int} \ S \), there exists an \( \varepsilon > 0 \) such that \( \{ z : \| z - x_2 \| < \varepsilon \} \subseteq S \). Let \( y \) be such that

\[
y = \lambda x_1 + (1 - \lambda)x_2, \tag{2.1}
\]

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

Since \( x_1 \in \text{cl} \ S \),

\[
\left\{ x : \| x - x_1 \| < \frac{(1 - \lambda)\varepsilon - \| z - y \|}{\lambda} \right\} \cap S
\]
is not empty. In particular, there exists a \( z_1 \in S \) such that
\[
\|z_1 - x_1\| < \frac{(1 - \lambda)\varepsilon - \|z - y\|}{\lambda}.
\]
(2.2)

Now let \( z_2 = \frac{z - \lambda z_1}{1 - \lambda} \). From (2.1), the Schwartz inequality, and (2.2), we get
\[
\|z_2 - x_2\| = \left\| \frac{z - \lambda z_1}{1 - \lambda} - x_2 \right\| = \left\| \frac{(z - \lambda z_1) - (y - \lambda x_1)}{1 - \lambda} \right\|
\]
\[
= \frac{1}{1 - \lambda} \left\| (z - y) + \lambda (x_1 - z_1) \right\|
\]
\[
\leq \frac{1}{1 - \lambda} (\|z - y\| + \lambda \|x_1 - z_1\|)
\]
\[
< \varepsilon.
\]

Therefore, \( z_2 \in S \). By the definition of \( z_2 \), note that \( z = \lambda z_1 + (1 - \lambda)z_2 \); and since both \( z_1 \) and \( z_2 \) belong to \( S \), \( z \) also belongs to \( S \). We have shown that any \( z \) with \( \|z - y\| < (1 - \lambda)\varepsilon \) belongs to \( S \). Therefore, \( y \in \text{int } S \) and the proof is complete.

**Corollary 1**

Let \( S \) be a convex set. Then \( \text{int } S \) is convex.

**Corollary 2**

Let \( S \) be a convex set with a nonempty interior. Then \( \text{cl } S \) is convex.

**Proof**

Let \( x_1, x_2 \in \text{cl } S \). Pick \( z \in \text{int } S \) (by assumption, \( \text{int } S \neq \emptyset \)). By the theorem, \( \lambda x_2 + (1 - \lambda)z \in \text{int } S \) for each \( \lambda \in (0,1) \). Now fix \( \mu \in (0,1) \). By the theorem, \( \mu x_1 + (1 - \mu)(\lambda x_2 + (1 - \lambda)z) \in \text{int } S \subset S \) for each \( \lambda \in (0,1) \). If we take the limit as \( \lambda \) approaches 1, it follows that \( \mu x_1 + (1 - \mu)x_2 \in \text{cl } S \), and the proof is complete.
Corollary 3

Let $S$ be a convex set with a nonempty interior. Then $\text{cl}(\text{int } S) = \text{cl } S$.

**Proof**

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

Corollary 4

Let $S$ be a convex set with a nonempty interior. Then $\text{int}(\text{cl } S) = \text{int } S$.

**Proof**

Note that $\text{int } S \subseteq \text{int } (\text{cl } S)$. Let $x_1 \in \text{int } (\text{cl } S)$. We need to show that $x_1 \in \text{int } S$. There exists an $\varepsilon > 0$ such that $\|y - x_1\| < \varepsilon$ implies that $y \in \text{cl } S$. Now let $x_2 \neq x_1$ belong to $\text{int } S$ and let $y = (1 + \Delta)x_1 - \Delta x_2$, where $\Delta = \varepsilon / (2\|x_1 - x_2\|)$. Since $\|y - x_1\| = \varepsilon / 2$, $y \in \text{cl } S$. But $x_1 = \lambda y + (1 - \lambda)x_2$, where $\lambda = 1 / (1 + \Delta) \in (0,1)$. Since $y \in \text{cl } S$ and $x_2 \in \text{int } S$, then, by the theorem, $x_1 \in \text{int } S$, and the proof is complete.

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

2.3 Weierstrass's Theorem

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

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


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

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

### 2.3.1 Theorem

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

**Proof**

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

---

**Figure 2.5** Nonexistence of a minimizing solution.
2.4 Separation and Support of Sets

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

**Minimum Distance from a Point to a Convex Set**

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

$$
\|a + b\|^2 = \|a\|^2 + \|b\|^2 + 2a^Tb
$$

$$
\|a - b\|^2 = \|a\|^2 + \|b\|^2 - 2a^Tb.
$$

By adding we get

$$
\|a + b\|^2 + \|a - b\|^2 = 2\|a\|^2 + 2\|b\|^2.
$$

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

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

**2.4.1 Theorem**

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

Figure 2.6 Parallelogram law.
Convex Sets

51

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

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

$$\|\bar{x} - \frac{\bar{x} + x'}{2}\| \leq \frac{1}{2}\|\bar{x} - x\| + \frac{1}{2}\|\bar{x} - x'\| = r.$$

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

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

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

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

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

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

Also,

$$\|y - \bar{x} - \lambda(x - \bar{x})\|^2 = \|y - \bar{x}\|^2 + \lambda^2\|x - \bar{x}\|^2 - 2\lambda(y - \bar{x})'(x - \bar{x}). \quad (2.4)$$

From (2.3) and (2.4) we get

$$2\lambda(y - \bar{x})'(x - \bar{x}) \leq \lambda^2\|x - \bar{x}\|^2 \quad (2.5)$$
for all \(0 \leq \lambda \leq 1\). Dividing (2.5) by any such \(\lambda > 0\) and letting \(\lambda \to 0^+\), the result follows.

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

### Hyperplanes and Separation of Two Sets

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

#### 2.4.2 Definition

A hyperplane \(H\) in \(\mathbb{R}^n\) is a collection of points of the form \(\{x : p^t x = \alpha\}\), where \(p\) is a nonzero vector in \(\mathbb{R}^n\) and \(\alpha\) is a scalar. The vector \(p\) is called the normal vector of the hyperplane. A hyperplane \(H\) defines two closed half-spaces \(H^+ = \{x : p^t x \geq \alpha\}\) and \(H^- = \{x : p^t x \leq \alpha\}\) and the two open half-spaces \(\{x : p^t x > \alpha\}\) and \(\{x : p^t x < \alpha\}\).

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

![Figure 2.7 Minimum distance to a closed convex set.](image-url)
As an example, consider $H = \{(x_1, x_2, x_3, x_4): x_1 + x_2 - x_3 + 2x_4 = 4\}$. The normal vector is $p = (1,1,-1,2)'$. Alternatively, the hyperplane can be written in reference to any point in $H$: for example, $\bar{x} = (0,6,0,-1)'$. In this case we write $H = \{(x_1, x_2, x_3, x_4): x_1 + (x_2 - 6) - x_3 + 2(x_4 + 1) = 0\}$.

**2.4.3 Definition**

Let $S_1$ and $S_2$ be nonempty sets in $\mathbb{R}^n$. A hyperplane $H = \{x: p'x = \alpha\}$ is said to separate $S_1$ and $S_2$ if $p'x \geq \alpha$ for each $x \in S_1$ and $p'x \leq \alpha$ for each $x \in S_2$. If, in addition, $S_1 \cup S_2 \subseteq H$, $H$ is said to properly separate $S_1$ and $S_2$. The hyperplane $H$ is said to strictly separate $S_1$ and $S_2$ if $p'x > \alpha$ for each $x \in S_1$ and $p'x < \alpha$ for each $x \in S_2$. The hyperplane $H$ is said to strongly separate $S_1$ and $S_2$ if $p'x \geq \alpha + \varepsilon$ for each $x \in S_1$ and $p'x \leq \alpha$ for each $x \in S_2$, where $\varepsilon$ is a positive scalar.

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

**Separation of a Convex Set and a Point**

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

**2.4.4 Theorem**

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

---

Figure 2.8 Hyperplane and corresponding half-spaces.
Figure 2.9 Various types of separation.

**Proof**

The set $S$ is a nonempty closed convex set and $y \not\in S$. Hence, by Theorem 2.4.1, there exists a unique minimizing point $\bar{x} \in S$ such that $(x - \bar{x})' (y - \bar{x}) \leq 0$ for each $x \in S$.

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

**Corollary 1**

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

**Proof**

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

**Corollary 2**

Let $S$ be a nonempty set, and let $y \not\in \text{cl } \text{conv}(S)$, the closure of the convex hull of $S$. Then there exists a strongly separating hyperplane for $S$ and $y$. 
The result follows by letting \( \text{cl} \, \text{conv}(S) \) play the role of \( S \) in Theorem 2.4.4. The following statements are equivalent to the conclusion of the theorem. The reader is asked to verify this equivalence. Note that statements 1 and 2 are equivalent only in this special case since \( y \) is a point. Note also that in Theorem 2.4.4, we have \( \alpha = p^T \bar{x} = \max \{ p^T x : x \in S \} \), since for any \( x \in S \), \( p^T (y - x) = (y - \bar{x})^T (\bar{x} - x) \geq 0 \).

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

**Farkas's Theorem as a Consequence of Theorem 2.4.4**

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

**System 1:** \( Ax \leq 0 \) and \( c^T x > 0 \) for some \( x \in \mathbb{R}^n \).

**System 2:** \( A'y = c \) and \( y \geq 0 \) for some \( y \in \mathbb{R}^m \).

If we denote the columns of \( A' \) by \( a_1, \ldots, a_m \), System 2 has a solution if \( c \) lies in the convex cone generated by \( a_1, \ldots, a_m \). System 1 has a solution if the closed convex cone \( \{ x : Ax \leq 0 \} \) and the open half-space \( \{ x : c^T x > 0 \} \) have a nonempty intersection. These two cases are illustrated geometrically in Figure 2.10.

**2.4.5 Theorem (Farkas’s Theorem)**

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

**System 1:** \( Ax \leq 0 \) and \( c^T x > 0 \) for some \( x \in \mathbb{R}^n \).

**System 2:** \( A'y = c \) and \( y \geq 0 \) for some \( y \in \mathbb{R}^m \).

**Proof**

Suppose that System 2 has a solution; that is, there exists \( y \geq 0 \) such that \( A'y = c \). Let \( x \) be such that \( Ax \leq 0 \). Then \( c^T x = y^T Ax \leq 0 \). Hence, System 1 has no solution. Now suppose that System 2 has no solution. Form the set \( S = \{ x : x = A'y, y \geq 0 \} \). Note that \( S \) is a closed convex set and that \( c \not\in S \). By Theorem
2.4.4, there exists a vector $p \in \mathbb{R}^n$ and a scalar $\alpha$ such that $p'c > \alpha$ and $p'x \leq \alpha$ for all $x \in S$. Since $0 \in S$, $\alpha \geq 0$, so $p'c > 0$. Also, $\alpha \geq p'A'y = y'Ap$ for all $y \geq 0$. Since $y \geq 0$ can be made arbitrarily large, the last inequality implies that $Ap \leq 0$. We have therefore constructed a vector $p \in \mathbb{R}^n$ such that $Ap \leq 0$ and $c'p > 0$. Hence, System 1 has a solution, and the proof is complete.

**Corollary 1 (Gordan’s Theorem)**

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

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

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

**Proof**

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

$$\begin{bmatrix} A' \\ e' \end{bmatrix} y = (0, \ldots, 0, 1)' \text{ and } y \geq 0 \text{ for some } y \in \mathbb{R}^m; \text{ that is, } A'y = 0, e'y = 1, \text{ and}$$
y ≥ 0 for some y ∈ R^m. This is equivalent to System 2 of the corollary, and hence the result follows.

**Corollary 2**

Let A be an m × n matrix and c be an n-vector. Then exactly one of the following two systems has a solution:

- **System 1:** Ax ≤ 0, x ≥ 0, c^T x > 0 for some x ∈ R^n.
- **System 2:** A^T y ≥ c, y ≥ 0 for some y ∈ R^m.

**Proof**

The result follows by writing the first set of constraints of System 2 as equalities and, accordingly, replacing A^T in the theorem by [A^T, -I].

**Corollary 3**

Let A be an m × n matrix, B be an ℓ × n matrix, and c be an n-vector. Then exactly one of the following two systems has a solution:

- **System 1:** Ax ≤ 0, Bx = 0, c^T x > 0 for some x ∈ R^n.
- **System 2:** A^T y + B^T z = c, y ≥ 0 for some y ∈ R^m and z ∈ R^ℓ.

**Proof**

The result follows by writing z = z_1 - z_2, where z_1 ≥ 0 and z_2 ≥ 0 in System 2 and, accordingly, replacing A^T in the theorem by [A^T, B^T, -B^T].

**Support of Sets at Boundary Points**

2.4.6 Definition

Let S be a nonempty set in R^n, and let x ∈ ∂S. A hyperplane H = {x: p^T (x - x̄) = 0} is called a supporting hyperplane of S at x̄ if either S ⊆ H⁺, that is, p^T (x - x̄) ≥ 0 for each x ∈ S, or else, S ⊆ H⁻, that is, p^T (x - x̄) ≤ 0 for each x ∈ S. If, in addition, S ⊆ H, H is called a proper supporting hyperplane of S at x̄.

Note that Definition 2.4.6 can be stated equivalently as follows. The hyperplane H = {x: p^T (x - x̄) = 0} is a supporting hyperplane of S at x̄ ∈ ∂S if p^T x̄ = inf{p^T x: x ∈ S} or else p^T x̄ = sup{p^T x: x ∈ S}. This follows by noting that either x̄ ∈ S, or if x̄ ∉ S, then since x̄ ∈ ∂S, there exist points in S arbitrarily
close to $\bar{x}$ and hence arbitrarily close in the value of the function $p'x$ to the value $p'\bar{x}$.

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

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

2.4.7 Theorem

Let $S$ be a nonempty convex set in $R^n$, and let $\bar{x} \in \partial S$. Then there exists a hyperplane that supports $S$ at $\bar{x}$; that is, there exists a nonzero vector $p$ such that

$$p'(x - \bar{x}) \leq 0$$

for each $x \in \text{cl } S$.

Proof

Since $\bar{x} \in \partial S$, there exists a sequence $\{y_k\}$ not in cl $S$ such that $y_k \to \bar{x}$. By Theorem 2.4.4, corresponding to each $y_k$ there exists a $p_k$ with norm 1 such that $p_k'y_k > p_k'x$ for each $x \in \text{cl } S$. (In Theorem 2.4.4, the normal vector can be normalized by dividing it by its norm, so that $\|p_k\| = 1$.) Since $\{p_k\}$ is bounded,
it has a convergent subsequence \( \{p_k\}_K \) with limit \( p \) whose norm is also equal to 1. Considering this subsequence, we have \( p_k y_k > p_k x \) for each \( x \in cl S \).

Fixing \( x \in cl S \) and taking limits as \( k \in \mathcal{K} \) approaches \( \infty \), we get \( p(x - \bar{x}) \leq 0 \). Since this is true for each \( x \in cl S \), the result follows.

**Corollary 1**

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

**Proof**

If \( \bar{x} \in cl S \), the corollary follows from Theorem 2.4.4. On the other hand, if \( \bar{x} \in \partial S \), the corollary reduces to Theorem 2.4.7.

**Corollary 2**

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

**Proof**

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

**Corollary 3**

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

**Proof**

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

**Separation of Two Convex Sets**

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

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

$$\inf \{ p'x : x \in S_1 \} \geq \sup \{ p'x : x \in S_2 \}. $$

Proof

Let $S = S_1 \cup S_2 = \{ x_1 - x_2 : x_1 \in S_1 \text{ and } x_2 \in S_2 \}$. Note that $S$ is a convex set. Furthermore, $0 \in S$, because otherwise $S_1 \cap S_2$ will be nonempty. By Corollary 1 of Theorem 2.4.7, there exists a nonzero $p \in \mathbb{R}^n$ such that $p'x \geq 0$ for all $x \in S$. This means that $p'x_1 \geq p'x_2$ for all $x_1 \in S_1$ and $x_2 \in S_2$, and the result follows.

Corollary 1

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

$$\inf \{ p'x : x \in S_1 \} \geq \sup \{ p'x : x \in S_2 \}. $$

Proof

Replace $S_2$ by $\text{int } S_2$, apply the theorem, and note that

$$\sup \{ p'x : x \in S_2 \} = \sup \{ p'x : x \in \text{int } S_2 \}. $$

Corollary 2

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

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

Gordan’s Theorem as a Consequence of Theorem 2.4.8

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

2.4.9 Theorem (Gordan’s Theorem)

Let \( A \) be an \( m \times n \) matrix. Then exactly one of the following systems has a solution:

- **System 1:** \( Ax < 0 \) for some \( x \in \mathbb{R}^n \).
- **System 2:** \( A'p = 0 \) and \( p \geq 0 \) for some nonzero \( p \in \mathbb{R}^m \).

**Proof**

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

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

\[ S_1 = \{ z : z = Ax, x \in \mathbb{R}^n \} \]
\[ S_2 = \{ z : z < 0 \}. \]

Note that \( S_1 \) and \( S_2 \) are nonempty convex sets such that \( S_1 \cap S_2 = \emptyset. \) Then, by Theorem 2.4.8, there exists a hyperplane that separates \( S_1 \) and \( S_2; \) that is, there exists a nonzero vector \( p \) such that

\[ p'Ax \geq p'z \quad \text{for each } x \in \mathbb{R}^n \text{ and } z \in \text{cl}S_2. \]

Since each component of \( z \) can be made an arbitrarily large negative number, we must have \( p \geq 0. \) Also, by letting \( z = 0, \) we must have \( p'Ax \geq 0 \) for each \( x \in \mathbb{R}^n. \)

By choosing \( x = -A'p, \) it follows that \( -\|A'p\|^2 \geq 0, \) and thus \( A'p = 0. \) Hence, System 2 has a solution, and the proof is complete.

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

2.4.10 Theorem (Strong Separation)

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

\[ \inf\{ p'x : x \in S_1 \} \geq \varepsilon + \sup\{ p'x : x \in S_2 \}. \]
Proof

Let $S = S_1 \ominus S_2$, and note that $S$ is a convex set and that $0 \not\in S$. We shall show that $S$ is closed. Let $\{x_k\}$ in $S$ converge to $x$. By the definition of $S$, $x_k = y_k - z_k$, where $y_k \in S_1$ and $z_k \in S_2$. Since $S_1$ is compact, there is a subsequence $\{y_k\}_k$ with limit $y$ in $S_1$. Since $y_k - z_k \to x$ and $y_k \to y$ for $k \to \infty$, $z_k \to z$ for $k \to \infty$. Since $S_2$ is closed, $z \in S_2$. Therefore, $x = y - z$ with $y \in S_1$ and $z \in S_2$. Therefore, $x \in S$ and hence $S$ is closed. By Theorem 2.4.4, there is a nonzero $p$ and an $\epsilon$ such that $p'x \geq \epsilon$ for each $x \in S$ and $p'0 < \epsilon$. Therefore, $\epsilon > 0$. By the definition of $S$, we conclude that $p'x_1 \geq \epsilon + p'x_2$ for each $x_1 \in S_1$ and $x_2 \in S_2$, and the result follows.

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

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

Corollary 1

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

2.5 Convex Cones and Polarity

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

Figure 2.13 Nonexistence of a strongly separating hyperplane.
2.5.1 Definition

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

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

2.5.2 Definition

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

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

2.5.3 Lemma

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

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

Figure 2.14 Cones.

Figure 2.15 Polar cones.
3. \( S_1 \subseteq S_2 \) implies that \( S_2^* \subseteq S_1^* \).

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

2.5.4 Theorem

Let \( C \) be a nonempty closed convex cone. Then \( C = C^{**} \).

Proof

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

Farkas’s Theorem as a Consequence of Theorem 2.5.4

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

System 1: \( Ax \leq 0 \) implies that \( c'x \leq 0 \).

System 2: \( A'y = c, y \geq 0 \).

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

System 1: \( Ax \leq 0, c'x > 0 \) (i.e., \( c \notin C^{**} = C \)).

System 2: \( A'y = c, y \geq 0 \) (i.e., \( c \in C \)).

2.6 Polyhedral Sets, Extreme Points, and Extreme Directions

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

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

2.6.1 Definition

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

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

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

Figure 2.16 illustrates the polyhedral set

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

Extreme Points and Extreme Directions

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

Figure 2.16 Polyhedral set.
2.6.2 Definition

Let $S$ be a nonempty convex set in $\mathbb{R}^n$. A vector $x \in S$ is called an extreme point of $S$ if $x = \lambda x_1 + (1 - \lambda) x_2$ with $x_1, x_2 \in S$, and $\lambda \in (0, 1)$ implies that $x = x_1 = x_2$.

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

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

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

2.6.3 Definition

Let $S$ be a nonempty, closed convex set in $\mathbb{R}^n$. A nonzero vector $d$ in $\mathbb{R}^n$ is called a direction, or a recession direction, of $S$ if for each $x \in S$, $x + \lambda d \in S$ for all $\lambda \geq 0$. Two directions $d_1$ and $d_2$ of $S$ are called distinct if $d_1 \neq \alpha d_2$ for any $\alpha > 0$. A
direction $d$ of $S$ is called an extreme direction if it cannot be written as a positive linear combination of two distinct directions; that is, if $d = \lambda_1 d_1 + \lambda_2 d_2$ for $\lambda_1, \lambda_2 > 0$, then $d_1 = \alpha d_2$ for some $\alpha > 0$.

To illustrate, consider $S = \{(x_1, x_2) : x_2 \geq \frac{1}{2}x_1\}$, shown in Figure 2.18. The directions of $S$ are nonzero vectors that make an angle less than or equal to 45° with the vector $(0,1)'$. In particular, $d_1 = (1,1)'$ and $d_2 = (-1,1)'$ are two extreme directions of $S$. Any other direction of $S$ can be represented as a positive linear combination of $d_1$ and $d_2$.

Characterization of Extreme Points and Extreme Directions for Polyhedral Sets

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

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

$$Bx_B + Nx_N = b \quad \text{and} \quad x_B \geq 0, x_N \geq 0.$$ 

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

2.6.4 Theorem (Characterization of Extreme Points)

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

Figure 2.18 Extreme directions.
where \( B \) is an \( m \times m \) invertible matrix satisfying \( B^{-1}b \geq 0 \). Any such solution is called a basic feasible solution (BFS) for \( S \).

**Proof**

Suppose that \( A \) can be decomposed into \([B, N]\) with \( x = \begin{bmatrix} B^{-1}b \\ 0 \end{bmatrix} \) and \( B^{-1}b \geq 0 \). It is obvious that \( x \in S \). Now suppose that \( x = \lambda x_1 + (1 - \lambda)x_2 \) with \( x_1, x_2 \in S \) for some \( \lambda \in (0,1) \). In particular, let \( x'_1 = (x_{11}, x_{12}) \) and \( x'_2 = (x_{21}, x_{22}) \). Then

\[
\begin{bmatrix} B^{-1}b \\ 0 \end{bmatrix} = \lambda \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} + (1 - \lambda) \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}.
\]

Since \( x_{12}, x_{22} \geq 0 \) and \( \lambda \in (0,1) \), it follows that \( x_{12} = x_{22} = 0 \). But this implies that \( x_{11} = x_{21} = B^{-1}b \) and hence \( x = x_1 = x_2 \). This shows that \( x \) is an extreme point of \( S \). Conversely, suppose that \( x \) is an extreme point of \( S \). Without loss of generality, suppose that \( x = (x_1,\ldots,x_k,0,\ldots,0)' \), where \( x_1,\ldots,x_k \) are positive. We shall first show that \( a_1,\ldots,a_k \) are linearly independent. By contradiction, suppose that there exist scalars \( \lambda_1,\ldots,\lambda_k \) not all zero such that \( \sum_{j=1}^k \lambda_j a_j = 0 \). Let \( \lambda = (\lambda_1,\ldots,\lambda_k,0,\ldots,0)' \). Construct the following two vectors, where \( \alpha > 0 \) is chosen such that \( x_1, x_2 \geq 0 \):

\[
x_1 = x + \alpha \lambda \quad \text{and} \quad x_2 = x - \alpha \lambda.
\]

Note that

\[
Ax_1 = Ax + \alpha A\lambda = Ax + \alpha \sum_{j=1}^k \lambda_j a_j = b,
\]

and similarly \( Ax_2 = b \). Therefore, \( x_1, x_2 \in S \), and since \( \alpha > 0 \) and \( \lambda \neq 0 \), \( x_1 \) and \( x_2 \) are distinct. Moreover, \( x = (1/2)x_1 + (1/2)x_2 \). This contradicts the fact that \( x \) is an extreme point. Thus, \( a_1,\ldots,a_k \) are linearly independent, and since \( A \) has rank \( m, m-k \) of the last \( n-k \) columns may be chosen such that they, together with the first \( k \) columns, form a linearly independent set of \( m \)-vectors. To simplify the notation, suppose that these columns are \( a_{k+1},\ldots,a_m \). Thus, \( A \) can be written as
Convex Sets

\[ A = [B, N], \text{ where } B = [a_1, \ldots, a_m] \text{ is of full rank. Furthermore, } B^{-1}b = (x_1, \ldots, x_k, 0, \ldots, 0)^T, \text{ and since } x_j > 0 \text{ for } j = 1, \ldots, k, \text{ } B^{-1}b \geq 0. \text{ This completes the proof.} \]

**Corollary**

The number of extreme points of \( S \) is finite.

**Proof**

The number of extreme points is less than or equal to

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

which is the maximum number of possible ways to choose \( m \) columns of \( A \) to form \( B \).

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

**2.6.5 Theorem (Existence of Extreme Points)**

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

**Proof**

Let \( x \in S \) and, without loss of generality, suppose that \( x = (x_1, \ldots, x_k, 0, \ldots, 0)^T \), where \( x_j > 0 \) for \( j = 1, \ldots, k \). If \( a_1, \ldots, a_k \) are linearly independent, \( k \leq m \) and \( x \) is an extreme point. Otherwise, there exist scalars \( \lambda_1, \ldots, \lambda_k \) with at least one positive component such that \( \sum_{j=1}^{k} \lambda_j a_j = 0 \). Define \( \alpha > 0 \) as follows:

\[
\alpha = \min_{1 \leq j \leq k} \left\{ \frac{x_j}{\lambda_j} : \lambda_j > 0 \right\} = \frac{x_j}{\lambda_j}, \text{ say.}
\]

Consider the point \( x' \) whose \( j \)th component \( x'_j \) is given by

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

Note that \( x'_j \geq 0 \) for \( j = 1, \ldots, k \) and \( x'_j = 0 \) for \( j = k+1, \ldots, n \). Moreover, \( x'_i = 0 \), and
Thus, so far, we have constructed a new point \( x' \) with at most \( k - 1 \) positive components. The process is continued until the positive components correspond to linearly independent columns, which results in an extreme point. Thus, we have shown that \( S \) has at least one extreme point, and the proof is complete.

**Extreme Directions**

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

\[
Ad = 0, \quad d \geq 0.
\]

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

**2.6.6 Theorem (Characterization of Extreme Directions)**

Let \( S = \{ x : Ax = b, x \geq 0 \} \neq \emptyset \), where \( A \) is an \( m \times n \) matrix of rank \( m \) and \( b \) is an \( m \)-vector. A vector \( \bar{d} \) is an extreme direction of \( S \) if and only if \( A \) can be decomposed into \([B, N]\) such that \( B^{-1}a_j \leq 0 \) for some column \( a_j \) of \( N \), and \( \bar{d} \) is a positive multiple of

\[
d = \begin{pmatrix} -B^{-1}a_j \\ e_j \end{pmatrix},
\]

where \( e_j \) is an \( n - m \) vector of zeros except for a 1 in position \( j \).

**Proof**

If \( B^{-1}a_j \leq 0, d \geq 0 \). Furthermore, \( Ad = 0 \), so that \( d \) is a direction of \( S \). We now show that \( d \) is indeed an extreme direction. Suppose that \( d = \lambda_1 d_1 + \lambda_2 d_2 \), where \( \lambda_1, \lambda_2 > 0 \) and \( d_1, d_2 \) are directions of \( S \). Noting that \( n - m - 1 \) components of \( d \) are equal to zero, the corresponding components of \( d_1 \) and \( d_2 \) must also be equal to zero. Thus, \( d_1 \) and \( d_2 \) could be written as follows:

\[
d_1 = \alpha_1 \begin{pmatrix} d_{11} \\ e_j \end{pmatrix}, \quad d_2 = \alpha_2 \begin{pmatrix} d_{21} \\ e_j \end{pmatrix},
\]

where \( \alpha_1, \alpha_2 > 0 \). Noting that \( Ad_1 = Ad_2 = 0 \), it can easily be verified that \( d_{11} = d_{21} = -B^{-1}a_j \). Thus, \( d_1 \) and \( d_2 \) are not distinct, which implies that \( d \) is an extreme direction. Since \( \bar{d} \) is a positive multiple of \( d \), it is also an extreme direction.

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

\[
\sum_{j=1}^{n} a_j x_j' = \sum_{j=1}^{k} a_j (x_j - \alpha_1 a_j) = \sum_{j=1}^{k} a_j x_j - \alpha \sum_{j=1}^{k} a_j a_j = b - 0 = b.
\]
\[ \vec{d} = (\vec{d}_1, \ldots, \vec{d}_k, 0, \ldots, \vec{d}_j, \ldots, 0)' , \]

where \( \vec{d}_i > 0 \) for \( i = 1, \ldots, k \) and for \( i = j \). We claim that \( a_1, \ldots, a_k \) are linearly independent. By contradiction, suppose that this were not the case. Then there would exist scalars \( \lambda_1, \ldots, \lambda_k \) not all zero such that \( \sum_{i=1}^k \lambda_i a_i = 0 \). Let \( \lambda = (\lambda_1, \ldots, \lambda_k, 0, \ldots, 0)' \) and choose \( \alpha > 0 \) sufficiently small such that both

\[ d_1 = \vec{d} + \alpha \lambda \quad \text{and} \quad d_2 = \vec{d} - \alpha \lambda \]

are nonnegative. Note that

\[ A d_1 = A \vec{d} + \alpha A \lambda = 0 + \alpha \sum_{i=1}^k a_i \lambda_i = 0. \]

Similarly, \( A d_2 = 0 \). Since \( d_1, d_2 \geq 0 \), they are both directions of \( S \). Note also that they are distinct, since \( \alpha > 0 \) and \( \lambda \neq 0 \). Furthermore, \( \vec{d} = (1/2)d_1 + (1/2)d_2 \), contradicting the assumption that \( \vec{d} \) is an extreme direction. Thus, \( a_1, \ldots, a_k \) are linearly independent, and since rank \( A \) is equal to \( m \), it is clear that \( k \leq m \). Then there must exist \( m - k \) vectors from among the set of vectors \( \{a_i : i = k+1, \ldots, n; \ i \neq j \} \) which, together with \( a_1, \ldots, a_k \), form a linearly independent set of \( m \)-vectors. Without loss of generality, suppose that these vectors are \( a_{k+1}, \ldots, a_m \).

Denote \( [a_1, \ldots, a_m] \) by \( B \), and note that \( B \) is invertible. Thus, \( 0 = A \vec{d} = B \hat{d} + a_j \vec{d}_j \), where \( \hat{d} \) is a vector of the first \( m \) components of \( \vec{d} \). Therefore, \( \hat{d} = -\vec{d}_j B^{-1} a_j \) and hence the vector \( \vec{d} \) is of the form \( \vec{d} = \vec{d}_j \begin{pmatrix} -B^{-1} a_j \\ e_j \end{pmatrix} \). Noting that \( \vec{d} \geq 0 \) and that \( \vec{d}_j > 0 \), \( B^{-1} a_j \leq 0 \), and the proof is complete.

**Corollary**

The number of extreme directions of \( S \) is finite.

**Proof**

For each choice of a matrix \( B \) from \( A \), there are \( n - m \) possible ways to extract the column \( a_j \) from \( N \). Therefore, the maximum number of extreme directions is bounded by

\[ \binom{n-m}{m} = \frac{n!}{m!(n-m-1)!}. \]
Chapter 2

Representation of Polyhedral Sets in Terms of Extreme Points and Extreme Directions

By definition, a polyhedral set is the intersection of a finite number of half-spaces. This representation may be thought of as an outer representation. A polyhedral set can also be described fully by an inner representation by means of its extreme points and extreme directions. This fact is fundamental to several linear and nonlinear programming procedures.

The main result can be stated as follows. Let $S$ be a nonempty polyhedral set of the form \( \{ x : Ax = b, \ x \geq 0 \} \). Then any point in $S$ can be represented as a convex combination of its extreme points plus a nonnegative linear combination of its extreme directions. Of course, if $S$ is bounded, it contains no directions, so any point in $S$ can be described as a convex combination of its extreme points.

In Theorem 2.6.7 it is assumed implicitly that the extreme points and extreme directions of $S$ are finite in number. This fact follows from the corollaries to Theorems 2.6.4 and 2.6.6. (See Exercises 2.30 to 2.32 for an alternative, constructive derivation of this theorem.)

2.6.7 Theorem (Representation Theorem)

Let $S$ be a nonempty polyhedral set in $\mathbb{R}^n$ of the form \( \{ x : Ax = b \text{ and } x \geq 0 \} \), where $A$ is an $m \times n$ matrix with rank $m$. Let $x_1, \ldots, x_k$ be the extreme points of $S$ and $d_1, \ldots, d_\ell$ be the extreme directions of $S$. Then \( x \in S \) if and only if $x$ can be written as

\[
x = \sum_{j=1}^{k} \lambda_j x_j + \sum_{j=1}^{\ell} \mu_j d_j
\]

(2.6)\[
\sum_{j=1}^{k} \lambda_j = 1
\]

(2.7)\[
\lambda_j \geq 0 \quad \text{for } j = 1, \ldots, k
\]

(2.8)\[
\mu_j \geq 0 \quad \text{for } j = 1, \ldots, \ell.
\]

Proof

Construct the following set:

\[
\Lambda = \left\{ \sum_{j=1}^{k} \lambda_j x_j + \sum_{j=1}^{\ell} \mu_j d_j : \sum_{j=1}^{k} \lambda_j = 1, \lambda_j \geq 0 \text{ for all } j, \mu_j \geq 0 \text{ for all } j \right\}.
\]

Note that $\Lambda$ is a closed convex set. Furthermore, by Theorem 2.6.5, $S$ has at least one extreme point and hence $\Lambda$ is not empty. Also note that $\Lambda \subseteq S$. To show that $S \subseteq \Lambda$, suppose by contradiction that there is a $z \in S$ such that $z \notin \Lambda$.

By Theorem 2.4.4 there exists a scalar $\alpha$ and a nonzero vector $p$ in $\mathbb{R}^n$ such that
\[ p'z > \alpha \quad (2.9) \]

where the \( \lambda_j \) values satisfy (2.6), (2.7), and (2.8). Since \( \mu_j \) can be made arbitrarily large, (2.9) holds true only if \( p'd_j \leq 0 \) for \( j = 1, \ldots, \ell \). From (2.9), by letting \( \mu_j = 0 \) for all \( j \), \( \lambda_j = 1 \), and \( \lambda_i = 0 \) for \( i \neq j \), it follows that \( p'x_j \leq \alpha \) for each \( j = 1, \ldots, k \). Since \( p'z > \alpha \), we have \( p'z > p'x_j \), for all \( j \). Summarizing, there exists a nonzero vector \( p \) such that

\[ p'z > p'x_j \quad \text{for } j = 1, \ldots, k \quad (2.10) \]

\[ p'd_j \leq 0 \quad \text{for } j = 1, \ldots, \ell. \quad (2.11) \]

Consider the extreme point \( \bar{x} \) defined as follows:

\[ p'\bar{x} = \max_{1 \leq j \leq k} p'x_j. \quad (2.12) \]

Since \( \bar{x} \) is an extreme point, by Theorem 2.6.4, \( \bar{x} = \begin{pmatrix} B^{-1}b \\ 0 \end{pmatrix} \), where \( A = [B, N] \) and \( B^{-1}b \geq 0 \). Without loss of generality assume that \( B^{-1}b > 0 \) (see Exercise 2.28). Since \( z \in S \), \( Az = b \) and \( z \geq 0 \). Therefore, \( Bz_B + Nz_N = b \) and hence \( z_B = B^{-1}b - B^{-1}Nz_N \), where \( z' \) is decomposed into \( (z_B', z_N') \). From (2.10) we have \( p'z - p'\bar{x} > 0 \), and decomposing \( p' \) into \( (p'_B, p'_N) \), we get

\[ 0 < p'z - p'\bar{x} = p'_B(B^{-1}b - B^{-1}Nz_N) + p'_Nz_N - p'_BB^{-1}b = (p'_N - p'_BB^{-1}N)z_N. \quad (2.13) \]

Since \( z_N \geq 0 \), from (2.13) it follows that there is a component \( j \geq m + 1 \) such that \( z_j > 0 \) and \( p_j - p'_BB^{-1}a_j > 0 \). We first show that \( y_j = B^{-1}a_j \leq 0 \). By contradiction, suppose that \( y_j \leq 0 \). Consider the vector \( d_j = \begin{bmatrix} -y_j \\ e_j \end{bmatrix} \), where \( e_j \) is an \( (n - m) \)-dimensional unit vector with 1 at position \( j \). By Theorem 2.6.6, \( d_j \) is an extreme direction of \( S \). From (2.11), \( p'd_j \leq 0 \), that is, \( -p'_BB^{-1}a_j + p_j \leq 0 \),
which contradicts the assertion that \( p_j - p_B'B^{-1}a_j > 0 \). Therefore, \( y_j \leq 0 \), and we can construct the following vector:

\[
x = \begin{pmatrix} \bar{b} \\ 0 \end{pmatrix} + \lambda \begin{pmatrix} -y_j \\ e_j \end{pmatrix},
\]

where \( \bar{b} \) is given by \( B^{-1}b \) and \( \lambda \) is given by

\[
\lambda = \min_{1 \leq i \leq m} \left\{ \frac{\bar{b}_i}{y_{ij}} : y_{ij} > 0 \right\} = \frac{\bar{b}_r}{y_{rj}} > 0.
\]

Note that \( x \geq 0 \) has, at most, \( m \) positive components, where the \( r \)th component drops to zero and the \( j \)th component is given by \( \lambda \). The vector \( x \) belongs to \( S \), since \( Ax = B(B^{-1}b - \lambda B^{-1}a_j) + \lambda a_j = B(B^{-1}b - \lambda B^{-1}a_j) + a_j = B' \). Since \( y_{rj} \neq 0 \), it can be shown that the vectors \( a_1, \ldots, a_{r-1}, a_{r+1}, \ldots, a_m, a_j \) are linearly independent. Therefore, by Theorem 2.6.4, \( x \) is an extreme point; that is, \( x \in \{x_1, x_2, \ldots, x_k\} \). Furthermore,

\[
p'x = (p'_B, p'_N) \begin{pmatrix} \bar{b} - \lambda y_j \\ \lambda e_j \end{pmatrix} = p'_B \bar{b} - \lambda p'_B y_j + \lambda p_j = p'_x + \lambda (p_j - p'_B B^{-1}a_j).
\]

Since \( \lambda > 0 \) and \( p_j - p'_B B^{-1}a_j > 0 \), \( p'x > p'_x \). Thus, we have constructed an extreme point \( x \) such that \( p'x > p'_x \), which contradicts (2.12). This contradiction asserts that \( z \) must belong to \( A \), and the proof is complete.

**Corollary (Existence of Extreme Directions)**

Let \( S \) be a nonempty polyhedral set of the form \( \{x : Ax = b, x \geq 0\} \) where \( A \) is an \( m \times n \) matrix with rank \( m \). Then \( S \) has at least one extreme direction if and only if it is unbounded.

**Proof**

If \( S \) has an extreme direction, it is obviously unbounded. Now suppose that \( S \) is unbounded and, by contradiction, suppose that \( S \) has no extreme directions. Using the theorem and the Schwartz inequality, it follows that

\[
\|x\| = \left\| \sum_{j=1}^{k} \lambda_j x_j \right\| \leq \sum_{j=1}^{k} \lambda_j \|x_j\| \leq \sum_{j=1}^{k} \|x_j\|
\]
for any $x \in S$. However, this violates the unboundedness assumption. Therefore, $S$ has at least one extreme direction, and the proof is complete.

### 2.7 Linear Programming and the Simplex Method

A linear programming problem is the minimization or the maximization of a linear function over a polyhedral set. Many problems can be formulated as, or approximated by, linear programs. Also, linear programming is often used in the process of solving nonlinear and discrete problems. In this section we describe the well-known simplex method for solving linear programming problems. The method is mainly based on exploiting the extreme points and directions of the polyhedral set defining the problem. Several other algorithms developed in this book can also be specialized to solve linear programming problems. In particular, Chapter 9 describes an efficient (polynomial-time) primal-dual, interior point, path-following algorithm, whose variants compete favorably with the simplex method.

Consider the following linear programming problem:

\[
\text{Minimize } \mathbf{c}' \mathbf{x} \\
\text{subject to } \mathbf{x} \in S,
\]

where $S$ is a polyhedral set in $\mathbb{R}^n$. The set $S$ is called the feasible region, and the linear function $\mathbf{c}' \mathbf{x}$ is called the objective function.

The optimum objective function value of a linear programming problem may be finite or unbounded. We give below a necessary and sufficient condition for a finite optimal solution. The importance of the concepts of extreme points and extreme directions in linear programming will be evident from the theorem.

#### 2.7.1 Theorem (Optimality Conditions in Linear Programming)

Consider the following linear programming problem: Minimize $\mathbf{c}' \mathbf{x}$ subject to $\mathbf{A} \mathbf{x} = \mathbf{b}$, $\mathbf{x} \geq \mathbf{0}$. Here $\mathbf{c}$ is an $n$-vector, $\mathbf{A}$ is an $m \times n$ matrix of rank $m$, and $\mathbf{b}$ is an $m$-vector. Suppose that the feasible region is not empty, and let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ be the extreme points and $\mathbf{d}_1, \ldots, \mathbf{d}_\ell$ be the extreme directions of the feasible region.

A necessary and sufficient condition for a finite optimal solution is that $\mathbf{c}' \mathbf{d}_j \geq 0$ for all $j = 1, \ldots, \ell$. If this condition holds true, there exists an extreme point $\mathbf{x}_i$ that solves the problem.

**Proof**

By Theorem 2.6.7, $\mathbf{A} \mathbf{x} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$ if and only if
\[
\mathbf{x} = \sum_{j=1}^{k} \lambda_j \mathbf{x}_j + \sum_{j=1}^{\ell} \mu_j \mathbf{d}_j
\]

\[
\sum_{j=1}^{k} \lambda_j = 1
\]

\[
\lambda_j \geq 0 \quad \text{for } j = 1, ..., k
\]

\[
\mu_j \geq 0 \quad \text{for } j = 1, ..., \ell.
\]

Therefore, the linear programming problem can be stated as follows:

Minimize \[
\mathbf{c}' \left( \sum_{j=1}^{k} \lambda_j \mathbf{x}_j + \sum_{j=1}^{\ell} \mu_j \mathbf{d}_j \right)
\]

subject to \[
\sum_{j=1}^{k} \lambda_j = 1
\]

\[
\lambda_j \geq 0 \quad \text{for } j = 1, ..., k
\]

\[
\mu_j \geq 0 \quad \text{for } j = 1, ..., \ell.
\]

Given feasibility, note that if \( \mathbf{c}' \mathbf{d}_j < 0 \) for some \( j \), \( \mu_j \) can be chosen arbitrarily large, leading to an unbounded optimal objective value. This shows that given feasibility, a necessary and sufficient condition for a finite optimum is \( \mathbf{c}' \mathbf{d}_j \geq 0 \) for \( j = 1, ..., \ell \). If this condition holds true, to minimize the objective function we may choose \( \mu_j = 0 \) for \( j = 1, ..., \ell \), and the problem reduces to minimizing \( \mathbf{c}' \left( \sum_{j=1}^{k} \lambda_j \mathbf{x}_j \right) \) subject to \( \sum_{j=1}^{k} \lambda_j = 1 \) and \( \lambda_j \geq 0 \) for \( j = 1, ..., k \). It is clear that the optimal solution to the latter problem is finite and found by letting \( \lambda_i = 1 \) and \( \lambda_j = 0 \) for \( j \neq i \), where the index \( i \) is given by \( \mathbf{c}' \mathbf{x}_i = \min_{1 \leq j \leq k} \mathbf{c}' \mathbf{x}_j \). Thus, there exists an optimal extreme point, and the proof is complete.

From Theorem 2.7.1, at least for the case in which the feasible region is bounded, one may be tempted to calculate \( \mathbf{c}' \mathbf{x}_j \) for \( j = 1, ..., k \) and then find \( \min_{1 \leq j \leq k} \mathbf{c}' \mathbf{x}_j \). Even though this is theoretically possible, it is computationally not advisable because the number of extreme points is usually prohibitively large.

**Simplex Method**

The simplex method is a systematic procedure for solving a linear programming problem by moving from extreme point to extreme point while improving (not worsening) the objective function value. This process continues until an optimal extreme point is reached and recognized, or else until an extreme direction \( \mathbf{d} \)
having $c'd < 0$ is found. In the latter case, we conclude that the objective value is unbounded, and we declare the problem to be "unbounded." Note that the unboundedness of the feasible region is a necessary but not sufficient condition for the problem to be unbounded.

Consider the following linear programming problem, in which the polyhedral set is defined in terms of equations and variables that are restricted to be nonnegative:

Minimize $c'x$

subject to $Ax = b$

$x \geq 0$.

Note that any polyhedral set can be put in the above standard format. For example, an inequality of the form $\sum_{j=1}^{n}a_{ij}x_{j} \leq b_{i}$ can be transformed into an equation by adding the nonnegative slack variable $s_{i}$, so that $\sum_{j=1}^{n}a_{ij}x_{j} + s_{i} = b_{i}$. Also, an unrestricted variable $x_{j}$ can be replaced by the difference of two nonnegative variables; that is, $x_{j} = x_{j}^{+} - x_{j}^{-}$, where $x_{j}^{+}, x_{j}^{-} \geq 0$. These and other manipulations could be used to put the problem in the above format. We shall assume for the time being that the constraint set admits at least one feasible point and that the rank of $A$ is equal to $m$.

By Theorem 2.7.1, at least in the case of a finite optimal solution, it suffices to concentrate on extreme points. Suppose that we have an extreme point $\bar{x}$. By Theorem 2.6.4, this point is characterized by a decomposition of $A$ into $[B, N]$, where $B = [a_{B_{1}}, \ldots, a_{B_{m}}]$ is an $m \times m$ matrix of full rank called the basis and $N$ is an $m \times (n - m)$ matrix. By Theorem 2.6.4, note that $\bar{x}$ could be written as $\bar{x} = (\bar{x}_{B}, \bar{x}_{N}) = (\bar{b}', 0')$, where $\bar{b} = B^{-1}b \geq 0$. The variables corresponding to the basis $B$ are called basic variables and are denoted by $x_{B_{1}}, \ldots, x_{B_{m}}$, whereas the variables corresponding to $N$ are called nonbasic variables.

Now let us consider a point $x$ satisfying $Ax = b$ and $x \geq 0$. Decompose $x'$ into $(x_{B}', x_{N}')$ and note that $x_{B}, x_{N} \geq 0$. Also, $Ax = b$ can be written as $Bx_{B} + Nx_{N} = b$. Hence,

$$x_{B} = B^{-1}b - B^{-1}Nx_{N}.$$  

Then, using (2.14) yields

$$c'x = c'_{B}x_{B} + c'_{N}x_{N}$$

$$= c'_{B}B^{-1}b + (c'_{N} - c'_{B}B^{-1}N)x_{N}$$  

$$= c'\bar{x} + (c'_{N} - c'_{B}B^{-1}N)x_{N}.$$  

(2.15)
Hence, if \( c'_N - c'_B B^{-1} N \geq 0 \), since \( x_N \geq 0 \), we have \( c'x \geq c'\bar{x} \) and that \( \bar{x} \) is an optimal extreme point. On the other hand, suppose that \( c'_N - c'_B B^{-1} N \not\geq 0 \). In particular, suppose that the \( j \)th component \( c_j - c'_B B^{-1} a_j \) is negative. Consider \( x = \bar{x} + \lambda d_j \), where

\[
d_j = \begin{pmatrix} -B^{-1}a_j \\ e_j \end{pmatrix}
\]

and where \( e_j \) is an \( n - m \) unit vector having a 1 at position \( j \). Then, from (2.15),

\[
c'x = c'\bar{x} + \lambda(c_j - c'_B B^{-1} a_j)
\]

and we get \( c'x < c'\bar{x} \) for \( \lambda > 0 \), since \( c_j - c'_B B^{-1} a_j < 0 \). We now consider the following two cases, where \( y_j = B^{-1}a_j \).

**Case 1:** \( y_j \leq 0 \). Note that \( A d_j = 0 \), and since \( A\bar{x} = b, A x = b \) for \( x = \bar{x} + \lambda d_j \) and for all values of \( \lambda \). Hence, \( x \) is feasible if and only if \( x \geq 0 \). This obviously holds true for all \( \lambda \geq 0 \) if \( y_j \leq 0 \). Thus, from (2.16), the objective function value is unbounded. In this case we have found an extreme direction \( d_j \) with \( c'd_j = c_j - c'_B B^{-1} a_j < 0 \) (see Theorems 2.7.1 and 2.6.6).

**Case 2:** \( y_j \not\leq 0 \). Let \( B^{-1}b = \bar{b} \), and let \( \lambda \) be defined by

\[
\lambda = \min_{1 \leq i \leq m} \left\{ \frac{\bar{b}_i}{y_{ij}} : y_{ij} > 0 \right\} = \frac{\bar{b}_r}{y_{rj}} \geq 0,
\]

where \( y_{ij} \) is the \( ij \)th component of \( y_j \). In this case, the components of \( x = \bar{x} + \lambda d_j \) are given by

\[
x_{B_i} = \frac{b_i - y_{rj} y_{ij}}{y_{rj}} \quad \text{for} \quad i = 1, \ldots, m
\]

\[
x_j = \frac{b_r}{y_{rj}}.
\]

and all other \( x_i \) values are equal to zero.
The positive components of $x$ can only be $x_{B_1}, \ldots, x_{B_{r-1}}, \ldots, x_{B_{r+1}}, \ldots, x_{B_m}$ and $x_j$. Hence, at most, $m$ components of $x$ are positive. It is easy to verify that their corresponding columns in $A$ are linearly independent. Therefore, by Theorem 2.6.4, the point $x$ is itself an extreme point. In this case we say that the basic variable $x_{B_r}$ left the basis and the nonbasic variable $x_j$ entered the basis in exchange.

Thus far, we have shown that given an extreme point, we can check its optimality and stop, or find an extreme direction leading to an unbounded solution, or find an extreme point having a better objective value (when $\lambda > 0$ in (2.17); else, only a revision of the basis matrix representing the current extreme point occurs). The process is then repeated.

**Summary of the Simplex Algorithm**

Outlined below is a summary of the simplex algorithm for a minimization problem of the form to minimize $c^T x$ subject to $Ax = b$, $x \geq 0$. A maximization problem can be either transformed into a minimization problem or else we have to modify Step 1 such that we stop if $c^T B^{-1} N - c^T_N \geq 0$ and introduce $x_j$ into the basis if $c_B^T B^{-1} a_j - c_j < 0$.

**Initialization Step** Find a starting extreme point $x$ with basis $B$. If such a point is not readily available, use artificial variables, as discussed later in the section.

**Main Step**

1. Let $x$ be an extreme point with basis $B$. Calculate $c_B^T B^{-1} N - c^T_N$. If this vector is nonpositive, stop; $x$ is an optimal extreme point. Otherwise, pick the most positive component $c_B^T B^{-1} a_j - c_j$. If $y_j = B^{-1} a_j \leq 0$, stop; the objective value is unbounded along the ray

$$\left\{ x + \lambda \begin{pmatrix} -y_j \\ e_j \end{pmatrix} : \lambda \geq 0 \right\},$$

where $e_j$ is a vector of zeros except for a 1 in position $j$. If, on the other hand, $y_j \not\leq 0$, go to Step 2.

2. Compute the index $r$ from (2.17) and form the new extreme point $x$ in (2.18). Form the new basis by deleting the column $a_{B_r}$ from $B$ and introducing $a_j$ in its place. Repeat Step 1.
Finite Convergence of the Simplex Method

If at each iteration, that is, one pass through the main step, we have $\vec{b} = B^{-1}b > 0$, then $\lambda$, defined by (2.17), would be strictly positive, and the objective value at the current extreme point would be strictly less than that at any of the previous iterations. This would imply that the current point is distinct from those generated previously. Since we have a finite number of extreme points, the simplex algorithm must stop in a finite number of iterations. If, on the other hand, $\vec{b}_r = 0$, then $\lambda = 0$, and we would remain at the same extreme point but with a different basis. In theory, this could happen an infinite number of times and may cause nonconvergence. This phenomenon, called cycling, sometimes occurs in practice. The problem of cycling can be overcome, but this topic is not discussed here. Most textbooks on linear programming give detailed procedures for avoiding cycling (see the Notes and References section at the end of this chapter).

Tableau Format of the Simplex Method

Suppose that we have a starting basis $B$ corresponding to an initial extreme point. The objective function and the constraints can be written as

- Objective row: $f - c_B^l x_B - c_N^l x_N = 0$
- Constraint rows: $B x_B + N x_N = b$

These equations can be displayed in the following simplex tableau, where the entries in the RHS column are the right-hand side constants of the equations.

<table>
<thead>
<tr>
<th>$f$</th>
<th>$x_B^l$</th>
<th>$x_N^l$</th>
<th>RHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$-c_B^l$</td>
<td>$-c_N^l$</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>$B$</td>
<td>$N$</td>
<td>$b$</td>
</tr>
</tbody>
</table>

The constraint rows are updated by multiplying by $B^{-1}$, and the objective row is updated by adding it to $c_B^l$ times the new constraint rows. We then get the following updated tableau. Note that the basic variables are indicated on the left-hand side and that $\vec{b} = B^{-1}b$.

<table>
<thead>
<tr>
<th>$f$</th>
<th>$x_B^l$</th>
<th>$x_N^l$</th>
<th>RHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$0^l$</td>
<td>$c_B^l B^{-1} N - c_N^l$</td>
<td>$c_B^l \vec{b}$</td>
</tr>
<tr>
<td>$0$</td>
<td>$I$</td>
<td>$B^{-1} N$</td>
<td>$\vec{b}$</td>
</tr>
</tbody>
</table>

Observe that the values of the basic variables and that of $f$ are recorded on the right-hand side of the tableau. Also, the vector $c_B^l B^{-1} N - c_N^l$ (the negative of
this vector is referred to as the vector of reduced cost coefficients) and the matrix \( \mathbf{B}^{-1} \mathbf{N} \) are stored conveniently under the nonbasic variables.

The above tableau displays all the information needed to perform Step 1 of the simplex method. If \( c'_B \mathbf{B}^{-1} \mathbf{N} - c'_N \leq 0 \), we stop; the current extreme point is optimal. Otherwise, upon examining the objective row, we can select a nonbasic variable having a positive value of \( c'_B \mathbf{B}^{-1} \mathbf{a}_j - c_j \). If \( \mathbf{B}^{-1} \mathbf{a}_j \leq 0 \), we stop; the problem is unbounded. Now suppose that \( y_j = \mathbf{B}^{-1} \mathbf{a}_j \leq 0 \). Since \( \mathbf{b} \) and \( y_j \) are recorded under RHS and \( x_j \), respectively, \( \lambda \) in (2.17) can easily be calculated from the tableau. The basic variable \( x_{B_r} \), corresponding to the minimum ratio of (2.17), leaves the basis and \( x_j \) enters the basis.

We would now like to update the tableau to reflect the new basis. This can be done by pivoting at the \( x_{B_r} \) row and the \( x_j \) column, that is, at \( y_{rj} \), as follows:

1. Divide the \( r \)th row corresponding to \( x_{B_r} \) by \( y_{rj} \).
2. Multiply the new \( r \)th row by \( y_{ij} \) and subtract it from the \( i \)th constraint row, for \( i = 1, \ldots, m, i \neq r \).
3. Multiply the new \( r \)th row by \( c'_B \mathbf{B}^{-1} \mathbf{a}_j - c_j \) and subtract it from the objective row.

The reader can easily verify that the above pivoting operation will update the tableau to reflect the new basis (see Exercise 2.37).

### 2.7.2 Example

Minimize \( x_1 - 3x_2 \)

subject to \( -x_1 + 2x_2 \leq 6 \)
\( x_1 + x_2 \leq 5 \)
\( x_1, x_2 \geq 0. \)

The problem is illustrated in Figure 2.19. It is clear that the optimal solution is \((4/3, 11/3)^f\) and that the corresponding value of the objective function is \(-29/3\).

To use the simplex method, we now introduce the two slack variables \( x_3 \geq 0 \) and \( x_4 \geq 0 \). This leads to the following standard format:

Minimize \( x_1 - 3x_2 \)

subject to \( -x_1 + 2x_2 + x_3 = 6 \)
\( x_1 + x_2 + x_4 = 5 \)
\( x_1, x_2, x_3, x_4 \geq 0. \)
Note that \( \mathbf{c} = (1, -3, 0, 0)' \), \( \mathbf{b} = \begin{pmatrix} 6 \\ 5 \end{pmatrix} \), and \( \mathbf{A} = \begin{bmatrix} -1 & 2 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix} \). By choosing \( \mathbf{B} = [\mathbf{a}_3, \mathbf{a}_4] \) = \( \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \), we note that \( \mathbf{B}^{-1}\mathbf{b} = \mathbf{b} \geq 0 \) and hence we have a starting basic feasible or extreme point solution. The corresponding tableau is:

<table>
<thead>
<tr>
<th>( f )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>RHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f )</td>
<td>1</td>
<td>-1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Note that \( x_2 \) enters and \( x_3 \) leaves the basis. The new basis is \( \mathbf{B} = [\mathbf{a}_2, \mathbf{a}_4] \).

<table>
<thead>
<tr>
<th>( f )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>RHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f )</td>
<td>1</td>
<td>1/2</td>
<td>0</td>
<td>-3/2</td>
<td>0</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0</td>
<td>-1/2</td>
<td>1</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>( \frac{3}{2} )</td>
<td>0</td>
<td>-1/2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Now \( x_1 \) enters and \( x_4 \) leaves the basis. The new basis is \( \mathbf{B} = [\mathbf{a}_2, \mathbf{a}_1] \).

<table>
<thead>
<tr>
<th>( f )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>RHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f )</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-4/3</td>
<td>-1/3</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1/3</td>
<td>1/3</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1/3</td>
<td>2/3</td>
</tr>
</tbody>
</table>

This solution is optimal since \( \mathbf{c}_B' \mathbf{B}^{-1} \mathbf{N} - \mathbf{c}_N' \leq 0 \). The three points corresponding to the three tableaux are shown in the \((x_1, x_2)\) space in Figure 2.19. We see that
the simplex method moved from one extreme point to another until the optimal solution was reached.

**Initial Extreme Point**

Recall that the simplex method starts with an initial extreme point. From Theorem 2.6.4, finding an initial extreme point of the set \( S = \{ x : Ax = b, x \geq 0 \} \) involves decomposing \( A \) into \( B \) and \( N \) with \( B^{-1}b \geq 0 \). In Example 2.7.2, an initial extreme point was available immediately. However, in many cases, an initial extreme point may not be conveniently available. This difficulty can be overcome by introducing *artificial variables*.

We discuss briefly two procedures for obtaining the initial extreme point. These are the two-phase method and the big-\( M \) method. For both methods, the problem is first put in the standard format \( Ax = b \) and \( x \geq 0 \), with the additional requirement that \( b \geq 0 \) (if \( b_i < 0 \), the \( i \)th constraint is multiplied by \(-1\)).

**Two-Phase Method** In this method the constraints of the problem are altered by the use of artificial variables so that an extreme point of the new system is at hand. In particular, the constraint system is modified to

\[
Ax + x_a = b \\
x, x_a \geq 0,
\]

where \( x_a \) is an artificial vector. Obviously, the solution \( x = 0 \) and \( x_a = b \) represents an extreme point of the above system. Since a feasible solution of the original system will be obtained only if \( x_a = 0 \), we can use the simplex method itself to minimize the sum of the artificial variables starting from the extreme point above. This leads to the following *Phase I* problem:

Minimize \( e^T x_a \)
subject to \( Ax + x_a = b \)
\( x, x_a \geq 0, \)

where \( e \) is a vector of ones. At the end of Phase I, either \( x_a \neq 0 \) or \( x_a = 0 \). In the former case we conclude that the original system is inconsistent; that is, the feasible region is empty. In the latter case, the artificial variables would drop from the basis and hence we would obtain an extreme point of the original system. Starting with this extreme point, *Phase II* of the simplex method minimizes the original objective \( c^T x \).

**Big-M Method** As in the two-phase method, the constraints are modified by the use of artificial variables so that an extreme point of the new

\footnote{It is possible that some of the artificial variables remain in the basis at the zero level at the end of Phase I. This case can easily be treated (see Bazarra et al. [2005]).}
system is immediately available. A large positive cost coefficient \( M \) is assigned to each artificial variable so that they will drop to zero level. This leads to the following problem:

Minimize \( c^T x + M e^T x_a \)
subject to \( Ax + x_a = b \)
\( x, x_a \geq 0 \).

We can execute the simplex method without actually specifying a numerical value for \( M \) by carrying the objective coefficients of \( M \) for the nonbasic variables as a separate vector. These coefficients identify precisely with the reduced objective coefficients for the Phase I problem, hence directly relating the two-phase and big-M methods. Consequently, we select nonbasic variables to enter that have a negative coefficient of \( M \) in the reduced cost vector (e.g., the most negative reduced cost), if any exist. When the coefficients of \( M \) in the reduced cost vector are all nonnegative, Phase I is complete. At this stage, if \( x_a = 0 \), we have a basic feasible solution to the original problem, and we can continue solving the original problem to termination (with an indication of unboundedness or optimality). On the other hand, if \( x_a \neq 0 \) at this stage, the optimal value of the Phase I problem is then positive, so we can conclude that the system \( Ax = b \) and \( x \geq 0 \) admits no feasible solutions.

### Duality in Linear Programming

The simplex method affords a simple derivation of a duality theorem for linear programming. Consider the linear program in standard form to minimize \( c^T x \) subject to \( Ax = b \) and \( x \geq 0 \). Let us refer to this as the primal problem \( P \). The following linear program is called the dual of the foregoing primal problem:

\[ \text{D: Maximize } b^T y \]
\[ \text{subject to } A^T y \leq c \]
\[ y \text{ unrestricted.} \]

We then have the following result that intimately relates the pair of linear programs \( P \) and \( D \) and permits the solution of one problem to be recovered from that of the other. As evident from the proof given below, this is possible via the simplex method, for example.

#### 2.7.3 Theorem

Let the pair of linear programs \( P \) and \( D \) be as defined above. Then we have

(a) **Weak duality result**: \( c^T x \geq b^T y \) for any feasible solution \( x \) to \( P \) and any feasible solution \( y \) to \( D \).

(b) **Unbounded-infeasible relationship**: If \( P \) is unbounded, \( D \) is infeasible, and vice versa.
Convex Sets

(c) **Strong duality result:** If both $P$ and $D$ are feasible, they both have optimal solutions with the same objective value.

**Proof**

For any pair of feasible solutions $x$ and $y$ to $P$ and $D$, respectively, we have $c'x \geq y'A = y'b$. This proves Part (a) of the theorem. Also, if $P$ is unbounded, $D$ must be infeasible, or else, a feasible solution to $D$ would provide a lower bound on the objective value for $P$ by Part (a). Similarly, if $D$ is unbounded, $P$ is infeasible, and this proves Part (b). Finally, suppose that both $P$ and $D$ are feasible. Then, by Part (b), neither could be unbounded, so they both have optimal solutions. In particular, let $\bar{x}' = (\bar{x}'_B, \bar{x}'_N)$ be an optimal basic feasible solution to $P$, where $\bar{x}_B = B^{-1}b$ and $\bar{x}_N = 0$ by Theorem 2.6.4. Now, consider the solution $\bar{y}' = c'_B B^{-1}$, where $c' = (c'_B, c'_N)$. We have $\bar{y}'A = c'_B B^{-1}[B, N] \leq [c'_B, c'_N]$, since $c'_B B^{-1}N - c'_N \leq 0$ by the optimality of the given basic feasible solution. Hence, $\bar{y}$ is feasible to $D$. Moreover, we have $\bar{y}'b = c'_B B^{-1}b = c'\bar{x}$; so, by Part (a), since $b'y \leq c'\bar{x}$ for all $y$ feasible to $D$, we have that $\bar{y}$ solves $D$ with the same optimal objective value as that for $P$. This completes the proof.

**Corollary 1**

If $D$ is infeasible, $P$ is unbounded or infeasible, and vice versa.

**Proof**

If $D$ is infeasible, $P$ could not have an optimal solution, or else, as in the proof of the theorem, we would be able to obtain an optimal solution for $D$, a contradiction. Hence, $P$ must be either infeasible or unbounded. Similarly, we can argue that if $P$ is infeasible, $D$ is either unbounded or infeasible.

**Corollary 2 (Farkas's Theorem as a Consequence of Theorem 2.7.3)**

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

- **System 1:** $Ax \leq 0$ and $c'x > 0$ for some $x \in \mathbb{R}^n$.
- **System 2:** $A'y = c$ and $y \geq 0$ for some $y \in \mathbb{R}^m$.

**Proof**

Consider the linear program $P$ to minimize $\{0'y : A'y = c, \ y \geq 0\}$. The dual $D$ to this problem is given by maximize $\{c'x : Ax \leq 0\}$. Then System 2 has no solution if and only if $P$ is infeasible, and this happens, by Part (a) of the theorem and Corollary 1, if and only if $D$ is unbounded, since $D$ is feasible. (For example, $x = 0$ is a feasible solution.) But because $Ax \leq 0$ defines a cone, this
happens in turn if and only if there exists an \( x \in R^n \) such that \( Ax \leq 0 \) and \( c'x > 0 \). Hence, System 2 has no solution if and only if System 1 has a solution, and this completes the proof.

**Corollary 3 (Complementary Slackness Conditions and Characterization of Optimality)**

Consider the pair of primal and dual linear programs \( P \) and \( D \) given above. Let \( \bar{x} \) be a primal feasible solution, and let \( \bar{y} \) be a dual feasible solution. Then \( \bar{x} \) and \( \bar{y} \) are, respectively, optimal to \( P \) and \( D \) if and only if \( \bar{v}, \bar{x} \geq 0 \), where

\[
\bar{v} = (\bar{v}_1, \bar{v}_2, ..., \bar{v}_n)' = c - A'\bar{y} \text{ is the vector of slack variables in the dual constraints for the dual solution } \bar{y}. \text{ (The latter conditions are called complementary slackness conditions, and when they hold true, the primal and dual solutions are called complementary slack solutions.)} \]

In particular, a given feasible solution is optimal for \( P \) if and only if there exists a complementary slack dual feasible solution, and vice versa.

**Proof**

Since \( \bar{x} \) and \( \bar{y} \) are, respectively, primal and dual feasible, we have \( Ax = b, \bar{x} \geq 0 \) and \( A'\bar{y} + \bar{v} = c, \bar{v} \geq 0 \), where \( \bar{v} \) is the vector of dual slack variables corresponding to \( \bar{y} \). Hence, \( c'\bar{x} - b'\bar{y} = (A'\bar{y} + \bar{v})'\bar{x} - (Ax)'\bar{y} = \bar{v}'\bar{x} \). By Theorem 2.7.3, the solutions \( \bar{x} \) and \( \bar{y} \) are, respectively, optimal to \( P \) and \( D \) if and only if \( c'\bar{x} = b'\bar{y} \). The foregoing statement asserts that this happens if and only if \( \bar{v}'\bar{x} = 0 \). But \( \bar{v} \geq 0 \) and \( \bar{x} \geq 0 \). Hence, \( \bar{v}'\bar{x} = 0 \) if and only if \( \bar{v}_j\bar{x}_j = 0 \) for all \( j = 1, ..., n \).

We have shown that \( \bar{x} \) and \( \bar{y} \) are, respectively, optimal to \( P \) and \( D \) if and only if the complementary slackness conditions hold true. The final statement of the theorem can now readily be verified using this result along with Theorem 2.7.3 and Corollary 1. This completes the proof.

**Exercises**

[2.1] Let \( S_1 \) and \( S_2 \) be nonempty sets in \( R^n \). Show that \( \text{conv}(S_1 \cap S_2) \subseteq \text{conv}(S_1) \cap \text{conv}(S_2) \). Is \( \text{conv}(S_1 \cap S_2) = \text{conv}(S_1) \cap \text{conv}(S_2) \) true in general? If not, give a counter-example.

[2.2] Let \( S \) be a polytope in \( R^n \). Show that \( S \) is a closed and bounded convex set.

[2.3] Let \( S \) be a closed set. Give an example to show that \( \text{conv}(S) \) is not necessarily closed. Specify a sufficient condition so that \( \text{conv}(S) \) is closed, and prove your assertion. (Hint: Suppose that \( S \) is compact.)
[2.4] Let $S$ be a polytope in $\mathbb{R}^n$, and let $S_j = \{\mu_j d_j : \mu_j \geq 0\}$, where $d_j$ is a non-zero vector in $\mathbb{R}^n$ for $j = 1, 2, \ldots, k$. Show that $S \oplus S_1 \oplus \cdots \oplus S_k$ is a closed convex set. (Note that Exercises 2.2 and 2.4 show that set $\Lambda$ in the proof of Theorem 2.6.7 is closed.)

[2.5] Identify the closure, interior, and boundary of each of the following convex sets:
   a. $S = \{x : x_1^2 + x_3^2 \leq x_2\}$.
   b. $S = \{x : 2 \leq x_1 \leq 5, x_2 = 4\}$.
   c. $S = \{x : x_1 + x_2 \leq 5, -x_1 + x_2 + x_3 \leq 7, x_1, x_2, x_3 \geq 0\}$.
   d. $S = \{x : x_1 + x_2 = 5, x_1 + x_2 + x_3 \leq 8\}$.
   e. $S = \{x : x_1^2 + x_2^2 + x_3^2 \leq 9, x_1 + x_3 = 2\}$.

[2.6] Let $S = \{x : x_1^2 + x_2^2 + x_3^2 \leq 4, x_1^2 - 4x_2 \leq 0\}$ and $y = (1, 0, 2)'$. Find the minimum distance from $y$ to $S$, the unique minimizing point, and a separating hyperplane.

[2.7] Let $S$ be a convex set in $\mathbb{R}^n$, $A$ be an $m \times n$ matrix, and $\alpha$ be a scalar. Show that the following two sets are convex.
   a. $AS = \{y : y = Ax, x \in S\}$.
   b. $\alpha S = \{\alpha x : x \in S\}$.

[2.8] Let $S_1 = \{x : x_1 = 0, 0 \leq x_2 \leq 1\}$ and $S_2 = \{x : 0 \leq x_1 \leq 1, x_2 = 2\}$. Describe $S_1 \oplus S_2$ and $S_1 \ominus S_2$.

[2.9] Prove Lemma 2.1.4.


[2.11] Let $S_1 = \{\lambda d_1 : \lambda \geq 0\}$ and $S_2 = \{\lambda d_2 : \lambda \geq 0\}$, where $d_1$ and $d_2$ are non-zero vectors in $\mathbb{R}^n$. Show that $S_1 \oplus S_2$ is a closed convex set.

[2.12] Let $S_1$ and $S_2$ be closed convex sets. Prove that $S_1 \oplus S_2$ is convex. Show by an example that $S_1 \oplus S_2$ is not necessarily closed. Prove that compactness of $S_1$ or $S_2$ is a sufficient condition for $S_1 \oplus S_2$ to be closed.

[2.13] Let $S$ be a nonempty set in $\mathbb{R}^n$. Show that $S$ is convex if and only if for each integer $k \geq 2$, the following holds true: $x_1, \ldots, x_k \in S$ implies that $\sum_{j=1}^{k} \lambda_j x_j \in S$, where $\sum_{j=1}^{k} \lambda_j = 1$ and $\lambda_j \geq 0$ for $j = 1, \ldots, k$.

[2.14] Let $C$ be a nonempty set in $\mathbb{R}^n$. Show that $C$ is a convex cone if and only if $x_1, x_2 \in C$ implies that $\lambda_1 x_1 + \lambda_2 x_2 \in C$ for all $\lambda_1, \lambda_2 \geq 0$. 
[2.15] Let $S_1 = \{x : A_1x \leq b_1\}$ and $S_2 = \{x : A_2x \leq b_2\}$ be nonempty. Define $S = S_1 \cup S_2$ and let $\hat{S} = \{x : x = y + z, A_1y \leq b_1\lambda_1, A_2z \leq b_2\lambda_2, \lambda_1 + \lambda_2 = 1, (\lambda_1, \lambda_2) \geq 0\}$.

a. Assuming that $S_1$ and $S_2$ are bounded, show that $\text{conv}(S) = \hat{S}$.

b. In general, show that $\text{cl conv}(S) = \hat{S}$.

[2.16] Let $S$ be a nonempty set in $\mathbb{R}^n$ and let $\bar{x} \in S$. Consider the set $C = \{y : y = \lambda(x - \bar{x}), \lambda \geq 0, x \in S\}$.

a. Show that $C$ is a cone and interpret it geometrically.

b. Show that $C$ is convex if $S$ is convex.

c. Suppose that $S$ is closed. Is it necessarily true that $C$ is closed? If not, under what conditions would $C$ be closed?

[2.17] Let $C_1$ and $C_2$ be convex cones in $\mathbb{R}^n$. Show that $C_1 \oplus C_2$ is also a convex cone and that $C_1 \oplus C_2 = \text{conv}(C_1 \cup C_2)$.

[2.18] Derive an explicit form of the polar $C^*$ of the following cones:

a. $C = \{(x_1, x_2) : 0 \leq x_2 \leq 2x_1\}$.

b. $C = \{(x_1, x_2) : x_2 \geq -3|x_1|\}$.

c. $C = \{x : x = Ap, p \geq 0\}$.

[2.19] Let $C$ be a nonempty convex cone in $\mathbb{R}^n$. Show that $C + C^* = \mathbb{R}^n$, that is, any point in $\mathbb{R}^n$ can be written as a point in the cone $C$ plus a point in its polar cone $C^*$. Is this representation unique? What if $C$ is a linear subspace?

[2.20] Let $S$ be a nonempty set in $\mathbb{R}^n$. The polar set of $S$, denoted by $S_p$, is given by $\{y : y'x \leq 1 \text{ for all } x \in S\}$.

a. Find the polar sets of the following two sets:

\[\{(x_1, x_2) : x_1^2 + x_2^2 \leq 4\}\] and \[\{(x_1, x_2) : 2x_1 + x_2 \leq 4, -2x_1 + x_2 \leq 2, x_1, x_2 \geq 0\}\]

b. Show that $S_p$ is a convex set. Is it necessarily closed?

c. If $S$ is a polyhedral set, is it necessarily true that $S_p$ is also a polyhedral set?

d. Show that if $S$ is a polyhedral set containing the origin, $S = S_{pp}$.

[2.21] Identify the extreme points and extreme directions of the following sets.

a. $S = \{x : 4x_2 \geq x_1^2, x_1 + 2x_2 + x_3 \leq 2\}$.

b. $S = \{x : x_1 + x_2 + 2x_3 \leq 4, x_1 + x_2 = 1, x_1, x_2, x_3 \geq 0\}$.

c. $S = \{x : x_2 \geq 2|x_1|, x_1^2 + x_2^2 \leq 2\}$.
[2.22] Establish the set of directions for each of the following convex sets.
   a. $S = \{ (x_1, x_2) : 4x_2 \geq x_1^2 \}$.
   b. $S = \{ (x_1, x_2) : x_1x_2 \geq 4, x_1 > 0 \}$.
   c. $S = \{ (x_1, x_2) : |x_1| + |x_2| \leq 2 \}$.

[2.23] Find the extreme points and directions of the following polyhedral sets.
   a. $S = \{ x : x_1 + 2x_2 + x_3 \leq 10, -x_1 + 3x_2 = 6, x_1, x_2, x_3 \geq 0 \}$.
   b. $S = \{ x : 2x_1 + 3x_2 \geq 6, x_1 - 2x_2 = 2, x_1, x_2 \geq 0 \}$.

[2.24] Consider the set $S = \{ x : -x_1 + 2x_2 \leq 4, x_1 - 3x_2 \leq 3, x_1, x_2 \geq 0 \}$. Identify all extreme points and extreme directions of $S$. Represent the point $(4,1)'$ as a convex combination of the extreme points plus a nonnegative combination of the extreme directions.

[2.25] Show that $C = \{ x : Ax \leq 0 \}$, where $A$ is an $m \times n$ matrix, has at most one extreme point, the origin.

[2.26] Let $S$ be a simplex in $R^n$ with vertices $x_1, x_2, ..., x_{k+1}$. Show that the extreme points of $S$ consist of its vertices.

[2.27] Let $S = \{ x : x_1 + 2x_2 \leq 4 \}$. Find the extreme points and directions of $S$. Can you represent any point in $S$ as a convex combination of its extreme points plus a nonnegative linear combination of its extreme directions? If not, discuss in relation to Theorem 2.6.7.

[2.28] Prove Theorem 2.6.7 if the nondegeneracy assumption $B^{-1}b > 0$ is dropped.

[2.29] Consider the nonempty unbounded polyhedral set $S = \{ x : Ax = b, x \geq 0 \}$, where $A$ is an $m \times n$ matrix of rank $m$. Starting with a direction of $S$, use the characterization of Theorem 2.6.6 to show how an extreme direction of $S$ can be constructed.

[2.30] Consider the polyhedral set $S = \{ x : Ax = b, x \geq 0 \}$, where $A$ is an $m \times n$ matrix of rank $m$. Then show that $\overline{x}$ is an extreme point of $S$ as defined by Theorem 2.6.4 if and only if there exist some $n$ linearly independent hyperplanes defining $S$ that are binding at $\overline{x}$.

[2.31] Consider the polyhedral set $S = \{ x : Ax = b, x \geq 0 \}$, where $A$ is an $m \times n$ matrix of rank $m$ and define $D = \{ d : Ad = 0, d \geq 0, e'd = 1 \}$, where $e$ is a vector of $n$ ones. Using the characterization of Theorem 2.6.6, show that $d \neq 0$ is an extreme direction of $S$ if and only if, when it is normalized to satisfy $e'd = 1$, it is an extreme point of $D$. Hence, show using Exercise 2.30 that the number of extreme directions is bounded above by $n!/(n - m - 1)!(m + 1)!$. Compare this with the corollary to Theorem 2.6.6.

[2.32] Let $S$ be a nonempty polyhedral set defined by $S = \{ x : Ax = b, x \geq 0 \}$, where $A$ is an $m \times n$ matrix of rank $m$. Consider any nonextreme point feasible solution $\overline{x} \in S$. 

Convex Sets
a. Show, using the definition of Exercise 2.30, that starting at $\bar{x}$, one can constructively recover an extreme point $\hat{x}$ of $S$ at which the hyperplanes binding at $\bar{x}$ are also binding at $\hat{x}$.

b. Assume that $S$ is bounded, and compute $\lambda_{\text{max}} = \max \{ \lambda : \hat{x} + \lambda (\bar{x} - \hat{x}) \in S \}$. Show that $\lambda_{\text{max}} > 0$ and that at the point $\bar{x} = \hat{x} + \lambda_{\text{max}} (\bar{x} - \hat{x})$, all the hyperplanes binding at $\bar{x}$ are also binding and that, in addition, at least one more linearly independent defining hyperplane of $S$ is binding.

c. Assuming that $S$ is bounded and noting how $\bar{x}$ is represented as a convex combination of the vertex $\hat{x}$ of $S$ and the point $\bar{x} \in S$ at which the number of linearly independent binding hyperplanes is, at least, one more than at $\bar{x}$, show how $\bar{x}$ can be represented constructively in terms of the extreme points of $S$.

d. Now, suppose that $S$ is unbounded. Define the nonempty, bounded polytope $\tilde{S} = \{ x \in S : e'x \leq M \}$, where $e$ is a vector of $n$ ones and $M$ is large enough so that any extreme point $\hat{x}$ of $S$ satisfies $e'\hat{x} < M$. Applying Part (c) to $S$ and simply using the definitions of extreme points and extreme directions as given in Exercises 2.30 and 2.31, prove the Representation Theorem 2.6.7.

[2.33] Let $S$ be a closed convex set in $\mathbb{R}^n$ and let $\bar{x} \in S$. Suppose that $d$ is a non-zero vector in $\mathbb{R}^n$ and that $\bar{x} + \lambda d \in S$ for all $\lambda \geq 0$. Show that $d$ is a direction of $S$.

[2.34] Solve the following problem by the simplex method:

Maximize $2x_1 + 3x_2 + 5x_3$
subject to $x_1 + 4x_2 - 2x_3 \leq 10$
$-x_1 + 2x_2 + 5x_3 \leq 3$
$x_1, x_2, x_3 \geq 0.$

What is an optimal dual solution to this problem?

[2.35] Consider the following problem:

Minimize $x_1 - 6x_2$
subject to $4x_1 + 3x_2 \leq 12$
$-x_1 + 2x_2 \leq 4$
$x_1 \leq 2.$

Find the optimal solution geometrically and verify its optimality by showing that $c'N - c'B^{-1}N \geq 0$. What is an optimal dual solution to this problem?

[2.36] Solve the following problem by the two-phase simplex method and by the big-$M$ method:
Maximize $-x_1 - 2x_2 + 2x_3$
subject to $2x_1 + 3x_2 + x_3 \geq 4$
\quad $x_1 + 2x_2 - x_3 \geq 6$
\quad $x_1 + 2x_3 \leq 12$
\quad $x_1, x_2, x_3 \geq 0.$

Also, identify an optimal dual solution to this problem from the final tableau obtained.

[2.37] Show in detail that pivoting at $y_{ij}$ updates the simplex tableau.

[2.38] Consider the following problem:

Minimize $c'x$
subject to $Ax = b$
\quad $x \geq 0,$

where $A$ is an $m \times n$ matrix of rank $m$. Let $x$ be an extreme point with corresponding basis $B$. Furthermore, suppose that $B^{-1}b > 0$. Use Farkas’s theorem to show that $x$ is an optimal point if and only if $c_N' - c_B' B^{-1} N \geq 0$.

[2.39] Consider the set $S = \{x : Ax = b, x \geq 0\}$, where $A$ is an $m \times n$ matrix and $b$ is an $m$-vector. Show that a nonzero vector $d$ is a direction of the set if and only if $Ad \leq 0$ and $d \geq 0$. Show how the simplex method can be used to generate such a direction.

[2.40] Consider the following problem:

Minimize $c'x$
subject to $Ax = b$
\quad $x \geq 0,$

where $A$ is an $m \times n$ matrix of rank $m$. Let $x$ be an extreme point with basis $B$, and let $\overline{b} = B^{-1}b$. Furthermore, suppose that $\overline{b}_i = 0$ for some component $i$. Is it possible that $x$ is an optimal solution even if $c_j - c_B' B^{-1} a_j < 0$ for some non-basic $x_j$? Discuss and give an example if this is possible.

[2.41] Let $P$: Minimize $\{c'x : Ax \geq b, x \geq 0\}$ and $D$: Maximize $\{b'y : A'y \leq c, \quad y \geq 0\}$. Show that $P$ and $D$ are a pair of primal and dual linear programs in the same sense as the pair of primal and dual programs of Theorem 2.7.3. (This pair is sometimes referred to as a symmetric pair of problems in canonical form.)

[2.42] Let $P$ and $D$ be a pair of primal and dual linear programs as in Theorem 2.7.3. Show that $P$ is infeasible if and only if the homogeneous version of $D$ (with right-hand sides replaced by zeros) is unbounded, and vice versa.
[2.43] Use Theorem 2.4.7 to construct an alternative proof for Theorem 2.2.2 by showing how the assumption that $\lambda x_1 + (1 - \lambda) x_2 \in \partial S$ leads readily to a contradiction.

[2.44] Let $A$ be an $m \times n$ matrix. Using Farkas’s theorem, prove that exactly one of the following two systems has a solution:

**System 1:** $Ax > 0$.

**System 2:** $A^t y = 0$, $y \geq 0$, $y \neq 0$.

(This is Gordan’s theorem developed in the text using Theorem 2.4.8.)

[2.45] Prove Gordan’s theorem 2.4.9 using the linear programming duality approach of Corollary 2 to Theorem 2.7.3.

[2.46] Prove that exactly one of the following two systems has a solution:

a. $Ax \geq 0$, $x \geq 0$, and $c^t x > 0$.

b. $A^t y \geq c$ and $y \leq 0$.

(*Hint: Use Farkas’s theorem.*)

[2.47] Show that the system $Ax \leq 0$ and $c^t x > 0$ has a solution $x$ in $R^3$, where $A = \begin{bmatrix} 2 & 1 & -3 \\ 2 & 2 & 0 \end{bmatrix}$ and $c = (-3, 1, -2)^t$.

[2.48] Let $A$ be a $p \times n$ matrix and $B$ be a $q \times n$ matrix. Show that if System 1 below has no solution, System 2 has a solution:

**System 1:** $Ax < 0$, $Bx = 0$ for some $x \in R^n$.

**System 2:** $A^t u + B^t v = 0$ for some nonzero $(u, v)$ with $u \geq 0$.

Furthermore, show that if $B$ has full rank, exactly one of the systems has a solution. Is this necessarily true if $B$ is not of full rank? Prove, or give a counterexample.

[2.49] Let $A$ be an $m \times n$ matrix and $c$ be an $n$-vector. Show that exactly one of the following two systems has a solution:

**System 1:** $Ax = c$.

**System 2:** $A^t y = 0$, $c^t y = 1$.

(This is a theorem of the alternative credited to Gale.)

[2.50] Let $A$ be a $p \times n$ matrix and $B$ be a $q \times n$ matrix. Show that exactly one of the following systems has a solution:

**System 1:** $Ax < 0$, $Bx = 0$ for some $x \in R^n$.

**System 2:** $A^t u + B^t v = 0$ for some nonzero $(u, v)$, $u \neq 0$, $u \geq 0$.

[2.51] Let $A$ be an $m \times n$ matrix. Show that the following two systems have solutions $\bar{x}$ and $\bar{y}$ such that $A\bar{x} + \bar{y} > 0$: 
System 1: \[ Ax \geq 0. \]
System 2: \[ A'y = 0, \ y \geq 0. \]

(This is an existence theorem credited to Tucker.)

[2.52] Let \( S_1 = \{ x : x_2 \geq e^{-x_1} \} \) and \( S_2 = \{ x : x_2 \leq -e^{-x_1} \} \). Show that \( S_1 \) and \( S_2 \) are disjoint convex sets and find a hyperplane that separates them. Does there exist a hyperplane that strongly separates \( S_1 \) and \( S_2 \)?

[2.53] Consider \( S = \{ x : x_1^2 + x_2^2 \leq 4 \} \). Represent \( S \) as the intersection of a collection of half-spaces. Find the half-spaces explicitly.

[2.54] Let \( S_1 \) and \( S_2 \) be convex sets in \( \mathbb{R}^n \). Show that there exists a hyperplane that strongly separates \( S_1 \) and \( S_2 \) if and only if

\[ \inf \{ \| x_1 - x_2 \| : x_1 \in S_1, \ x_2 \in S_2 \} > 0. \]

[2.55] Let \( S_1 \) and \( S_2 \) be nonempty, disjoint convex sets in \( \mathbb{R}^n \). Prove that there exist two nonzero vectors \( p_1 \) and \( p_2 \) such that

\[ p_1^T x_1 + p_2^T x_2 \geq 0 \quad \text{for all } x_1 \in S_1 \text{ and all } x_2 \in S_2. \]

Can you generalize this result for three or more disjoint convex sets?

[2.56] Let \( C_\varepsilon = \{ y : y = \lambda(x - \bar{x}), \ \lambda \geq 0, \ x \in S \cap N_\varepsilon(\bar{x}) \} \), where \( N_\varepsilon(\bar{x}) \) is an \( \varepsilon \)-neighborhood around \( \bar{x} \). Let \( T \) be the intersection of all such cones; that is, \( T = \cap \{ C_\varepsilon : \varepsilon > 0 \} \). Interpret the cone \( T \) geometrically. (\( T \) is called the cone of tangents of \( S \) at \( \bar{x} \) and is discussed in more detail in Chapter 5.)

[2.57] A linear subspace \( L \) of \( \mathbb{R}^n \) is a subset of \( \mathbb{R}^n \) such that \( x_1, x_2 \in L \) implies that \( \lambda_1 x_1 + \lambda_2 x_2 \in L \) for all scalars \( \lambda_1 \) and \( \lambda_2 \). The orthogonal complement \( L^\perp \) is defined by \( L^\perp = \{ y : y^T x = 0 \text{ for all } x \in L \} \). Show that any vector \( x \) in \( \mathbb{R}^n \) could be represented uniquely as \( x_1 + x_2 \), where \( x_1 \in L \) and \( x_2 \in L^\perp \). Illustrate by writing the vector \( (1,2,3)^T \) as the sum of two vectors in \( L \) and \( L^\perp \), respectively, where \( L = \{(x_1, x_2, x_3) : 2x_1 + x_2 - x_3 = 0 \} \).

Notes and References

In this chapter we treat the topic of convex sets. This subject was first studied systematically by Minkowski [1911], whose work contains the essence of the important results in this area. The topic of convexity is fully developed in a variety of good texts, and the interested reader may refer to Eggleston [1958], Rockafellar [1970], Stoer and Witzgall [1970], and Valentine [1964] for a more detailed analysis of convex sets.
In Section 2.1 we present some basic definitions and develop the Carathéodory theorem, which states that each point in the convex hull of any given set can be represented as the convex combination of \( n + 1 \) points in the set. This result can be sharpened by using the notion of the dimension of the set. Using this notion, several Carathéodory-type theorems can be developed. See, for example, Bazaraa and Shetty [1976], Eggleston [1958], and Rockafellar [1970].

In Section 2.2 we develop some topological properties of convex sets related to interior and closure points. Exercise 2.15 gives an important result due to Balas [1974] (see also Sherali and Shetty [1980c]) that is used to construct algebraically the closure convex hull for disjunctive program. In Section 2.3 we present an important theorem due to Weierstrass that is used widely to establish the existence of optimal solutions.

In Section 2.4 we present various types of theorems that separate disjoint convex sets. Support and separation theorems are of special importance in the area of optimization and are also widely used in game theory, functional analysis, and optimal control theory. An interesting application is the use of these results in coloring problems in graph theory. For further reading on support and separation of convex sets, see Eggleston [1958], Klee [1969], Mangasarian [1969a], Rockafellar [1970], Stoer and Witzgall [1970], and Valentine [1964].

Many of the results in Sections 2.2 and 2.4 can be strengthened by using the notion of relative interior. For example, every nonempty convex set has a nonempty relative interior. Furthermore, a hyperplane that properly separates two convex sets exists provided that they have disjoint relative interiors. Also, Theorem 2.2.2 and several of its corollaries can be sharpened using this concept. For a good discussion of relative interiors, see Eggleston [1958], Rockafellar [1970], and Valentine [1964].

In Section 2.5, a brief introduction to polar cones is given. For more details, see Rockafellar [1970]. In Section 2.6 we treat the important special case of polyhedral sets and prove the representation theorem, which states that every point in the set can be represented as a convex combination of the extreme points plus a nonnegative linear combination of the extreme directions. This result was first provided by Motzkin [1936] using a different approach. The representation theorem is also true for closed convex sets that contain no lines. For a proof of this result, see Bazaraa and Shetty [1976] and Rockafellar [1970]. An exhaustive treatment of convex polytopes is given by Grünbaum [1967]. Akgül [1988] and Sherali [1987b] provide geometrically motivated constructive proofs for the representation theorem based on definitions of extreme points and directions (see Exercises 2.30 to 2.32).

In Section 2.7 we present the simplex algorithm for solving linear programming problems. The simplex algorithm was developed by Dantzig in 1947. The efficiency of the simplex algorithm, advances in computer technology, and ability of linear programming to model large and complex problems led to the popularity of the simplex method and linear programming. The presentation of the simplex method in Section 2.7 is a natural extension of the material in Section 2.6 on polyhedral sets. For a further study of linear programming, see
Convex and concave functions have many special and important properties. For example, any local minimum of a convex function over a convex set is also a global minimum. In this chapter we introduce the important topics of convex and concave functions and develop some of their properties. As we shall learn in this and later chapters, these properties can be utilized in developing suitable optimality conditions and computational schemes for optimization problems that involve convex and concave functions.

Following is an outline of the chapter.

Section 3.1: Definitions and Basic Properties  We introduce convex and concave functions and develop some of their basic properties. Continuity of convex functions is proved, and the concept of a directional derivative is introduced.

Section 3.2: Subgradients of Convex Functions  A convex function has a convex epigraph and hence has a supporting hyperplane. This leads to the important notion of a subgradient of a convex function.

Section 3.3: Differentiable Convex Functions  In this section we give some characterizations of differentiable convex functions. These are helpful tools for checking convexity of simple differentiable functions.

Section 3.4: Minima and Maxima of Convex Functions  This section is important, since it deals with the questions of minimizing and maximizing a convex function over a convex set. A necessary and sufficient condition for a minimum is developed, and we provide a characterization for the set of alternative optimal solutions. We also show that the maximum occurs at an extreme point. This fact is particularly important if the convex set is polyhedral.

Section 3.5: Generalizations of Convex Functions  Various relaxations of convexity and concavity are possible. We present quasiconvex and pseudoconvex functions and develop some of their properties. We then discuss various types of convexity at a point. These types of convexity are sometimes sufficient for optimality, as shown in Chapter 4. (This section can be omitted by beginning readers, and later references to generalized convexity properties can largely be substituted simply by convexity.)
3.1 Definitions and Basic Properties

In this section we deal with some basic properties of convex and concave functions. In particular, we investigate their continuity and differentiability properties.

### 3.1.1 Definition

Let $f : S \to \mathbb{R}$, where $S$ is a nonempty convex set in $\mathbb{R}^n$. The function $f$ is said to be **convex** on $S$ if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

for each $x_1, x_2 \in S$ and for each $\lambda \in (0, 1)$. The function $f$ is called **strictly convex** on $S$ if the above inequality is true as a strict inequality for each distinct $x_1$ and $x_2$ in $S$ and for each $\lambda \in (0, 1)$. The function $f : S \to \mathbb{R}$ is called **concave** (strictly concave) on $S$ if $-f$ is convex (strictly convex) on $S$.

Now let us consider the geometric interpretation of convex and concave functions. Let $x_1$ and $x_2$ be two distinct points in the domain of $f$, and consider the point $\lambda x_1 + (1-\lambda)x_2$, with $\lambda \in (0, 1)$. Note that $\lambda f(x_1) + (1-\lambda)f(x_2)$ gives the weighted average of $f(x_1)$ and $f(x_2)$, while $f(\lambda x_1 + (1-\lambda)x_2)$ gives the value of $f$ at the point $\lambda x_1 + (1-\lambda)x_2$. So for a convex function $f$, the value of $f$ at points on the line segment $\lambda x_1 + (1-\lambda)x_2$ is less than or equal to the height of the chord joining the points $[x_1, f(x_1)]$ and $[x_2, f(x_2)]$. For a concave function, the chord is (on or) below the function itself. Hence, a function is both convex and concave if and only if it is **affine**. Figure 3.1 shows some examples of convex and concave functions.

The following are some examples of convex functions. By taking the negatives of these functions, we get some examples of concave functions.

1. $f(x) = 3x + 4$.
2. $f(x) = |x|$.
3. $f(x) = x^2 - 2x$.
4. $f(x) = -x^{1/2}$ if $x \geq 0$.
5. $f(x_1, x_2) = 2x_1^2 + x_2^2 - 2x_1x_2$.
6. $f(x_1, x_2, x_3) = x_1^4 + 2x_2^2 + 3x_3^2 - 4x_1 - 4x_2x_3$.

Note that in each of the above examples, except for Example 4, the function $f$ is convex over $\mathbb{R}^n$. In Example 4 the function is not defined for $x < 0$. One can readily construct examples of functions that are convex over a region but not over $\mathbb{R}^n$. For instance, $f(x) = x^3$ is not convex over $\mathbb{R}$ but is convex over $S = \{x : x \geq 0\}$.
Convex functions and Generalizations

The examples above cite some arbitrary illustrative instances of convex functions. In contrast, we give below some particularly important instances of convex functions that arise very often in practice and that are useful to remember.

1. Let $f_1, f_2, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ be convex functions. Then:
   
   (a) $f(x) = \sum_{j=1}^{k} \alpha_j f_j(x)$, where $\alpha_j > 0$ for $j = 1, 2, \ldots, k$ is a convex function (see Exercise 3.8).
   
   (b) $f(x) = \max\{f_1(x), f_2(x), \ldots, f_k(x)\}$ is a convex function (see Exercise 3.9).

2. Suppose that $g : \mathbb{R}^n \to \mathbb{R}$ is a concave function. Let $S = \{x : g(x) > 0\}$, and define $f : S \to \mathbb{R}$ as $f(x) = 1/g(x)$. Then $f$ is convex over $S$ (see Exercise 3.11).

3. Let $g : \mathbb{R} \to \mathbb{R}$ be a nondecreasing, univariate, convex function, and let $h : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then the composite function $f : \mathbb{R}^n \to \mathbb{R}$ defined as $f(x) = g[h(x)]$ is a convex function (see Exercise 3.10).

4. Let $g : \mathbb{R}^m \to \mathbb{R}$ be a convex function, and let $h : \mathbb{R}^n \to \mathbb{R}^m$ be an affine function of the form $h(x) = Ax + b$, where $A$ is an $m \times n$ matrix and $b$ is an $m \times 1$ vector. Then the composite function $f : \mathbb{R}^n \to \mathbb{R}$ defined as $f(x) = g[h(x)]$ is a convex function (see Exercise 3.16).

From now on, we concentrate on convex functions. Results for concave functions can be obtained easily by noting that $f$ is concave if and only if $-f$ is convex.

Associated with a convex function $f$ is the set $S_\alpha = \{x \in S : f(x) \leq \alpha\}$, $\alpha \in \mathbb{R}$, usually referred to as a level set. Sometimes this set is called a lower-level set, to differentiate it from the upper-level set $\{x \in S : f(x) \geq \alpha\}$, which has
properties similar to these for concave functions. Lemma 3.1.2 shows that $S_\alpha$ is convex for each real number $\alpha$. Hence, if $g_i: \mathbb{R}^n \to \mathbb{R}$ is convex for $i = 1, \ldots, m$, the set \{x : g_i(x) \leq 0, i = 1, \ldots, m\} is a convex set.

3.1.2 Lemma

Let $S$ be a nonempty convex set in $\mathbb{R}^n$, and let $f: S \to \mathbb{R}$ be a convex function. Then the level set $S_\alpha = \{x \in S : f(x) \leq \alpha\}$, where $\alpha$ is a real number, is a convex set.

Proof

Let $x_1, x_2 \in S_\alpha$. Thus, $x_1, x_2 \in S$ and $f(x_1) \leq \alpha$ and $f(x_2) \leq \alpha$. Now let $\lambda \in (0, 1)$ and $x = \lambda x_1 + (1 - \lambda)x_2$. By the convexity of $S$, we have that $x \in S$. Furthermore, by the convexity of $f$,

$$f(x) = \lambda f(x_1) + (1 - \lambda)f(x_2) \leq \lambda\alpha + (1 - \lambda)\alpha = \alpha.$$  

Hence, $x \in S_\alpha$, and therefore, $S_\alpha$ is convex.

Continuity of Convex Functions

An important property of convex and concave functions is that they are continuous on the interior of their domain. This fact is proved below.

3.1.3 Theorem

Let $S$ be a nonempty convex set in $\mathbb{R}^n$, and let $f: S \to \mathbb{R}$ be convex. Then $f$ is continuous on the interior of $S$.

Proof

Let $\bar{x} \in \text{int } S$. To prove continuity of $f$ at $\bar{x}$, we need to show that given $\varepsilon > 0$, there exists a $\delta > 0$ such that $\|x - \bar{x}\| \leq \delta$ implies that $|f(x) - f(\bar{x})| \leq \varepsilon$. Since $\bar{x} \in \text{int } S$, there exists a $\delta' > 0$ such that $\|x - \bar{x}\| \leq \delta'$ implies that $x \in S$. Construct $\theta$ as follows.

$$\theta = \max \{\max_{1 \leq i \leq n} \|f(x + \delta' e_i) - f(\bar{x}) - f(\bar{x} - \delta' e_i) - f(\bar{x})\|, \alpha\},$$

(3.1)

where $e_i$ is a vector of zeros except for a 1 at the $i$th position. Note that $0 \leq \theta < \infty$. Let

$$\delta = \min \left(\frac{\delta'}{n}, \frac{\varepsilon \theta}{n\theta}\right).$$

(3.2)
Choose an $x$ with $\|x - \bar{x}\| \leq \delta$. If $x_i - \bar{x}_i \geq 0$, let $z_i = \delta^i e_i$; otherwise, let $z_i = -\delta^i e_i$. Then $x - \bar{x} = \sum_{i=1}^{n} \alpha_i z_i$, where $\alpha_i \geq 0$ for $i = 1, \ldots, n$. Furthermore,

$$
\|x - \bar{x}\| = \delta^i \left( \sum_{i=1}^{n} \alpha_i^2 \right)^{1/2}.
$$

From (3.2), and since $\|x - \bar{x}\| \leq \delta$, it follows that $\alpha_i \leq 1/n$ for $i = 1, \ldots, n$. Hence, by the convexity of $f$, and since $0 \leq n\alpha_i \leq 1$, we get

$$
f(x) = f\left(\bar{x} + \sum_{i=1}^{n} \alpha_i z_i\right) = f\left[\frac{1}{n} \sum_{i=1}^{n} (\bar{x} + n\alpha_i z_i)\right]
\leq \frac{1}{n} \sum_{i=1}^{n} f(\bar{x} + n\alpha_i z_i)
= \frac{1}{n} \sum_{i=1}^{n} f(1 - n\alpha_i)\bar{x} + n\alpha_i(\bar{x} + z_i))
\leq \frac{1}{n} \sum_{i=1}^{n} [(1 - n\alpha_i) f(\bar{x}) + n\alpha_i f(\bar{x} + z_i)].
$$

Therefore, $f(x) - f(\bar{x}) \leq \sum_{i=1}^{n} \alpha_i [f(\bar{x} + z_i) - f(\bar{x})]$. From (3.1) it is obvious that $f(\bar{x} + z_i) - f(\bar{x}) \leq \theta$ for each $i$; and since $\alpha_i \geq 0$, it follows that

$$
f(x) - f(\bar{x}) \leq \theta \sum_{i=1}^{n} \alpha_i. \tag{3.4}
$$

Noting (3.3) and (3.2), it follows that $\alpha_i \leq \varepsilon/n\theta$, and (3.4) implies that $f(x) - f(\bar{x}) \leq \varepsilon$. So far, we have shown that $\|x - \bar{x}\| \leq \delta$ implies that $f(x) - f(\bar{x}) \leq \varepsilon$. By definition, this establishes the upper semicontinuity of $f$ at $\bar{x}$. To complete the proof, we need to establish the lower semicontinuity of $f$ at $\bar{x}$ as well, that is, to show that $f(\bar{x}) - f(x) \leq \varepsilon$. Let $y = 2\bar{x} - x$ and note that $\|y - \bar{x}\| \leq \delta$. Therefore, as above,

$$
f(y) - f(\bar{x}) \leq \varepsilon. \tag{3.5}
$$

But $\bar{x} = (1/2)y + (1/2)x$, and by the convexity of $f$, we have

$$
f(\bar{x}) \leq (1/2)f(y) + (1/2)f(x). \tag{3.6}
$$

Combining (3.5) and (3.6) above, it follows that $f(\bar{x}) - f(x) \leq \varepsilon$, and the proof is complete.
Note that convex and concave functions may not be continuous everywhere. However, by Theorem 3.1.3, points of discontinuity only are allowed at the boundary of $S$, as illustrated by the following convex function defined on $S = \{ x : -1 \leq x \leq 1 \}$:

$$f(x) = \begin{cases} 
    x^2 & \text{for } |x| < 1 \\
    2 & \text{for } |x| = 1.
\end{cases}$$

### Directional Derivative of Convex Functions

The concept of directional derivatives is particularly useful in the motivation and development of some optimality criteria and computational procedures in nonlinear programming, where one is interested in finding a direction along which the function decreases or increases.

#### 3.1.4 Definition

Let $S$ be a nonempty set in $\mathbb{R}^n$, and let $f: S \to R$. Let $\bar{x} \in S$ and $d$ be a nonzero vector such that $\bar{x} + \lambda d \in S$ for $\lambda > 0$ and sufficiently small. The directional derivative of $f$ at $\bar{x}$ along the vector $d$, denoted by $f'(\bar{x};d)$, is given by the following limit if it exists:

$$f'(\bar{x};d) = \lim_{\lambda \to 0^+} \frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda}.$$ 

In particular, the limit in Definition 3.1.4 exists for globally defined convex and concave functions as shown below. As evident from the proof of the following lemma, if $f: S \to R$ is convex on $S$, the limit exists if $\bar{x} \in \text{int } S$, but might be $-\infty$ if $\bar{x} \in \partial S$, even if $f$ is continuous at $\bar{x}$, as seen in Figure 3.2.

![Figure 3.2 Nonexistence of the directional derivative of $f$ at $\bar{x}$ in the direction $d$.](image-url)
3.1.5 Lemma

Let $f: \mathbb{R}^n \to \mathbb{R}$ be a convex function. Consider any point $\bar{x} \in \mathbb{R}^n$ and a nonzero direction $d \in \mathbb{R}^n$. Then the directional derivative $f'(\bar{x}; d)$, of $f$ at $\bar{x}$ in the direction $d$, exists.

**Proof**

Let $\lambda_2 > \lambda_1 > 0$. Noting the convexity of $f$, we have

$$f(\bar{x} + \lambda_1 d) = f\left[\lambda_1 \frac{\bar{x}}{\lambda_2} + \frac{\lambda_1}{\lambda_2} \bar{x} + \left(1 - \frac{\lambda_1}{\lambda_2}\right) d\right]$$

$$\leq \frac{\lambda_1}{\lambda_2} f(\bar{x} + \lambda_2 d) + \left(1 - \frac{\lambda_1}{\lambda_2}\right) f(\bar{x}).$$

This inequality implies that

$$\frac{f(\bar{x} + \lambda_1 d) - f(\bar{x})}{\lambda_1} \leq \frac{f(\bar{x} + \lambda_2 d) - f(\bar{x})}{\lambda_2}.$$ 

Thus, the difference quotient $[f(\bar{x} + \lambda d) - f(\bar{x})]/\lambda$ is monotone decreasing (nonincreasing) as $\lambda \to 0^+$. Now, given any $\lambda \geq 0$, we also have, by the convexity of $f$, that

$$f(\bar{x}) = f\left[\frac{\lambda}{1 + \lambda} (\bar{x} - d) + \frac{1}{1 + \lambda} (\bar{x} + \lambda d)\right]$$

$$\leq \frac{\lambda}{1 + \lambda} f(\bar{x} - d) + \frac{1}{1 + \lambda} f(\bar{x} + \lambda d).$$

So

$$\frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda} \geq f(\bar{x}) - f(\bar{x} - d).$$

Hence, the monotone decreasing sequence of values $[f(\bar{x} + \lambda d) - f(\bar{x})]/\lambda$, as $\lambda \to 0^+$, is bounded from below by the constant $f(\bar{x}) - f(\bar{x} - d)$. Hence, the limit in the theorem exists and is given by

$$\lim_{\lambda \to 0^+} \frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda} = \inf_{\lambda > 0} \frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda}.$$

### 3.2 Subgradients of Convex Functions

In this section, we introduce the important concept of subgradients of convex and concave functions via supporting hyperplanes to the epigraphs of convex functions and to the hypographs of concave functions.
Epigraph and Hypograph of a Function

A function \( f \) on \( S \) can be fully described by the set \( \{[x, f(x)] : x \in S\} \subset \mathbb{R}^{n+1} \), which is referred to as the graph of the function. One can construct two sets that are related to the graph of \( f \): the epigraph, which consists of points above the graph of \( f \); and the hypograph, which consists of points below the graph of \( f \). These notions are clarified in Definition 3.2.1.

3.2.1 Definition

Let \( S \) be a nonempty set in \( \mathbb{R}^n \), and let \( f : S \to \mathbb{R} \). The epigraph of \( f \), denoted by \( \text{epi} f \), is a subset of \( \mathbb{R}^{n+1} \) defined by
\[ \{ (x, y) : x \in S, \ y \in \mathbb{R}, \ y \geq f(x) \} . \]

The hypograph of \( f \), denoted by \( \text{hyp} f \), is a subset of \( \mathbb{R}^{n+1} \) defined by
\[ \{ (x, y) : x \in S, \ y \in \mathbb{R}, \ y \leq f(x) \} . \]

Figure 3.3 illustrates the epigraphs and hypographs of several functions. In Figure 3.3a, neither the epigraph nor the hypograph of \( f \) is a convex set. But in Figure 3.3b and c, respectively, the epigraph and hypograph of \( f \) are convex sets. It turns out that a function is convex if and only if its epigraph is a convex set and, equivalently, that a function is concave if and only if its hypograph is a convex set.

3.2.2 Theorem

Let \( S \) be a nonempty convex set in \( \mathbb{R}^n \), and let \( f : S \to \mathbb{R} \). Then \( f \) is convex if and only if \( \text{epi} f \) is a convex set.

\[ \text{Figure 3.3 Epigraphs and hypographs.} \]
Proof

Assume that \( f \) is convex, and let \((x_1, y_1)\) and \((x_2, y_2)\) \( \in \text{epi} f \); that is, \( x_1, x_2 \in S, \) \( y_1 \geq f(x_1), \) and \( y_2 \geq f(x_2). \) Let \( \lambda \in (0, 1). \) Then

\[
\lambda y_1 + (1 - \lambda) y_2 \geq \lambda f(x_1) + (1 - \lambda) f(x_2) \geq f(\lambda x_1 + (1 - \lambda) x_2),
\]

where the last inequality follows by the convexity of \( f. \) Note that \( \lambda x_1 + (1 - \lambda) x_2 \in S. \) Thus, \([\lambda x_1 + (1 - \lambda) x_2, \lambda y_1 + (1 - \lambda) y_2] \in \text{epi} f,\) and hence \( \text{epi} f \) is convex. Conversely, assume that \( \text{epi} f \) is convex, and let \( x_1, x_2 \in S. \) Then \([x_1, f(x_1)]\) and \([x_2, f(x_2)]\) belong to \( \text{epi} f,\) and by the convexity of \( \text{epi} f,\) we must have

\[
[\lambda x_1 + (1 - \lambda) x_2, \lambda f(x_1) + (1 - \lambda) f(x_2)] \in \text{epi} f \quad \text{for } \lambda \in (0, 1).
\]

In other words, \( \lambda f(x_1) + (1 - \lambda) f(x_2) \geq f[\lambda x_1 + (1 - \lambda) x_2] \) for each \( \lambda \in (0, 1);\) that is, \( f \) is convex. This completes the proof.

Theorem 3.2.2 can be used to verify the convexity or concavity of a given function \( f. \) Making use of this result, it is clear that the functions illustrated in Figure 3.3 are (a) neither convex nor concave, (b) convex, and (c) concave.

Since the epigraph of a convex function and the hypograph of a concave function are convex sets, they have supporting hyperplanes at points of their boundary. These supporting hyperplanes lead to the notion of subgradients, which is defined below.

3.2.3 Definition

Let \( S \) be a nonempty convex set in \( \mathbb{R}^n, \) and let \( f: S \rightarrow R \) be convex. Then \( \xi \) is called a subgradient of \( f \) at \( \bar{x} \in S \) if

\[
f(x) \geq f(\bar{x}) + \xi^T (x - \bar{x}) \quad \text{for all } x \in S.
\]

Similarly, let \( f: S \rightarrow R \) be concave. Then \( \xi \) is called a subgradient of \( f \) at \( \bar{x} \in S \) if

\[
f(x) \leq f(\bar{x}) + \xi^T (x - \bar{x}) \quad \text{for all } x \in S.
\]

From Definition 3.2.3 it follows immediately that the collection of subgradients of \( f \) at \( \bar{x} \) (known as the subdifferential of \( f \) at \( \bar{x} \)) is a convex set. Figure 3.4 shows examples of subgradients of convex and concave functions.

From the figure we see that the function \( f(\bar{x}) + \xi^T (x - \bar{x}) \) corresponds to a supporting hyperplane of the epigraph or the hypograph of the function \( f. \) The subgradient vector \( \xi \) corresponds to the slope of the supporting hyperplane.
3.2.4 Example

Let \( f(x) = \min \{ f_1(x), f_2(x) \} \), where \( f_1 \) and \( f_2 \) are as defined below:

\[
\begin{align*}
f_1(x) &= 4 - |x|, \quad x \in \mathbb{R} \\
f_2(x) &= 4 - (x - 2)^2, \quad x \in \mathbb{R}.
\end{align*}
\]

Since \( f_2(x) \geq f_1(x) \) for \( 1 \leq x \leq 4 \), \( f \) can be represented as follows:

\[
f(x) = \begin{cases} 
4 - x, & 1 \leq x \leq 4 \\
4 - (x - 2)^2, & \text{otherwise}.
\end{cases}
\]

In Figure 3.5 the concave function \( f \) is shown in dark lines. Note that \( \xi = -1 \) is the slope and hence the subgradient of \( f \) at any point \( x \) in the open interval \((1, 4)\). If \( x < 1 \) or \( x > 4 \), \( \xi = -2(x - 2) \) is the unique subgradient of \( f \). At the points \( x = 1 \) and \( x = 4 \), the subgradients are not unique because many supporting hyperplanes exist. At \( x = 1 \), the family of subgradients is characterized by \( \lambda \nabla f_1(1) + (1 - \lambda) \nabla f_2(1) = \lambda (-1) + (1 - \lambda) (2) = 2 - 3\lambda \) for \( \lambda \in [0, 1] \). In other words, any \( \xi \) in the interval \([-1, 2]\) is a subgradient of \( f \) at \( x = 1 \), and this corresponds to the slopes of the family of supporting hyperplanes of \( f \) at \( x = 1 \). At \( x = 4 \), the family of subgradients is characterized by \( \lambda \nabla f_1(4) + (1 - \lambda) \nabla f_2(4) = \lambda (-1) + (1 - \lambda) (-4) = -4 + 3\lambda \) for \( \lambda \in [0, 1] \). In other words, any \( \xi \) in the interval \([-4, -1]\) is a subgradient of \( f \) at \( x = 4 \). Exercise 3.27 addresses the general characterization of subgradients of functions of the form \( f(x) = \min \{ f_1(x), f_2(x) \} \).

The following theorem shows that every convex or concave function has at least one subgradient at points in the interior of its domain. The proof relies on the fact that a convex set has a supporting hyperplane at points of the boundary.
3.2.5 Theorem

Let $S$ be a nonempty convex set in $R^n$, and let $f: S \rightarrow R$ be convex. Then for $\bar{x} \in \text{int } S$, there exists a vector $\xi$ such that the hyperplane

$$H = \{(x, y) : y = f(\bar{x}) + \xi^T (x - \bar{x})\}$$

supports $\text{epi } f$ at $[\bar{x}, f(\bar{x})]$. In particular,

$$f(x) \geq f(\bar{x}) + \xi^T (x - \bar{x}) \quad \text{for each } x \in S;$$

that is, $\xi$ is a subgradient of $f$ at $\bar{x}$.

**Proof**

By Theorem 3.2.2, $\text{epi } f$ is convex. Noting that $[\bar{x}, f(\bar{x})]$ belongs to the boundary of $\text{epi } f$, by Theorem 2.4.7 there exists a nonzero vector $(\xi_0, \mu) \in R^n \times R$ such that

$$\xi_0^T (x - \bar{x}) + \mu [y - f(\bar{x})] \leq 0 \quad \text{for all } (x, y) \in \text{epi } f. \tag{3.7}$$

Note that $\mu$ is not positive, because otherwise, inequality (3.7) will be contradicted by choosing $y$ sufficiently large. We now show that $\mu < 0$. By contradiction, suppose that $\mu = 0$. Then $\xi_0^T (x - \bar{x}) \leq 0$ for all $x \in S$. Since $\bar{x} \in \text{int } S$, there exists a $\lambda > 0$ such that $\bar{x} + \lambda \xi_0 \in S$ and hence $\lambda \xi_0^T \xi_0 \leq 0$. This implies that $\xi_0 = 0$ and $(\xi_0, \mu) = (0, 0)$, contradicting the fact that $(\xi_0, \mu)$ is a nonzero vector. Therefore, $\mu < 0$. Denoting $\xi_0 / |\mu|$ by $\xi$ and dividing the inequality in (3.7) by $|\mu|$, we get

$$\xi^T (x - \bar{x}) - y + f(\bar{x}) \leq 0 \quad \text{for all } (x, y) \in \text{epi } f. \tag{3.8}$$
In particular, the hyperplane \( H = \{(x, y) : \ y = f(\bar{x}) + \xi^T(x - \bar{x})\} \) supports \( \text{epi} \ f \) at \([\bar{x}, f(\bar{x})]\). By letting \( y = f(\bar{x}) \) in (3.8), we get \( f(x) \geq f(\bar{x}) + \xi^T(x - \bar{x}) \) for all \( x \in S \), and the proof is complete.

**Corollary**

Let \( S \) be a nonempty convex set in \( R^n \), and let \( f : S \to R \) be strictly convex. Then for \( \bar{x} \in \text{int} \ S \) there exists a vector \( \xi \) such that

\[
f(x) > f(\bar{x}) + \xi^T(x - \bar{x}) \quad \text{for all } x \in S, x \neq \bar{x}.
\]

**Proof**

By Theorem 3.2.5 there exists a vector \( \xi \) such that

\[
f(x) \geq f(\bar{x}) + \xi^T(x - \bar{x}) \quad \text{for all } x \in S. \tag{3.9}
\]

By contradiction, suppose that there is an \( \hat{x} \neq \bar{x} \) such that \( f(\hat{x}) = f(\bar{x}) + \xi^T(\hat{x} - \bar{x}) \). Then, by the strict convexity of \( f \) for \( \lambda \in (0, 1) \), we get

\[
f[\lambda \bar{x} + (1 - \lambda)\hat{x}] < \lambda f(\bar{x}) + (1 - \lambda) f(\hat{x}) = f(\bar{x}) + (1 - \lambda) f(\bar{x}) + (1 - \lambda)\xi^T(\hat{x} - \bar{x}). \tag{3.10}
\]

But letting \( x = \lambda \bar{x} + (1 - \lambda)\hat{x} \) in (3.9), we must have

\[
f[\lambda \bar{x} + (1 - \lambda)\hat{x}] \geq f(\bar{x}) + (1 - \lambda)\xi^T(\hat{x} - \bar{x}),
\]

contradicting (3.10). This proves the corollary.

The converse of Theorem 3.2.5 is not true in general. In other words, if corresponding to each point \( \bar{x} \in \text{int} \ S \) there is a subgradient of \( f \), then \( f \) is not necessarily a convex function. To illustrate, consider the following example, where \( f \) is defined on \( S = \{(x_1, x_2) : 0 \leq x_1, x_2 \leq 1\} \):

\[
f(x_1, x_2) = \begin{cases} 
0, & 0 \leq x_1 \leq 1, \quad 0 < x_2 \leq 1 \\
\frac{1}{4} \left( x_1 - \frac{1}{2} \right)^2, & 0 \leq x_1 \leq 1, \quad x_2 = 0.
\end{cases}
\]

For each point in the interior of the domain, the zero vector is a subgradient of \( f \). However, \( f \) is not convex on \( S \) since \( \text{epi} \ f \) is clearly not a convex set. However, as the following theorem shows, \( f \) is indeed convex on \( \text{int} \ S \).

**3.2.6 Theorem**

Let \( S \) be a nonempty convex set in \( R^n \), and let \( f : S \to R \). Suppose that for each point \( \bar{x} \in \text{int} \ S \) there exists a subgradient vector \( \xi \) such that
Convex Functions and Generalizations

\[ f(x) \geq f(\bar{x}) + \xi'(x - \bar{x}) \quad \text{for each } x \in S. \]

Then, \( f \) is convex on \( \text{int} \ S \).

**Proof**

Let \( x_1, x_2 \in \text{int} \ S \), and let \( \lambda \in (0, 1) \). By Corollary 1 to Theorem 2.2.2, \( \text{int} \ S \) is convex, and we must have \( \lambda x_1 + (1 - \lambda)x_2 \in \text{int} \ S \). By assumption, there exists a subgradient \( \xi \) of \( f \) at \( \lambda x_1 + (1 - \lambda)x_2 \). In particular, the following two inequalities hold true:

\[ f(x_1) \geq f[\lambda x_1 + (1 - \lambda)x_2] + (1 - \lambda)\xi'(x_1 - x_2) \]
\[ f(x_2) \geq f[\lambda x_1 + (1 - \lambda)x_2] + \lambda\xi'(x_2 - x_1). \]

Multiplying the above two inequalities by \( \lambda \) and \( (1 - \lambda) \), respectively, and adding, we obtain

\[ \lambda f(x_1) + (1 - \lambda)f(x_2) \geq f[\lambda x_1 + (1 - \lambda)x_2], \]

and the result follows.

3.3 Differentiable Convex Functions

We now focus on differentiable convex and concave functions. First, consider the following definition of differentiability.

3.3.1 Definition

Let \( S \) be a nonempty set in \( \mathbb{R}^n \), and let \( f: S \to \mathbb{R} \). Then \( f \) is said to be **differentiable** at \( \bar{x} \in \text{int} \ S \) if there exist a vector \( \nabla f(\bar{x}) \), called the **gradient vector**, and a function \( \alpha: \mathbb{R}^n \to \mathbb{R} \) such that

\[ f(x) = f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \|x - \bar{x}\|\alpha(\bar{x}; x - \bar{x}) \quad \text{for each } x \in S, \]

where \( \lim_{x \to \bar{x}} \alpha(\bar{x}; x - \bar{x}) = 0 \). The function \( f \) is said to be differentiable on the open set \( S' \subseteq S \) if it is differentiable at each point in \( S' \). The representation of \( f \) above is called a **first-order (Taylor series) expansion** of \( f \) at (or about) the point \( \bar{x} \); and without the implicitly defined remainder term involving the function \( \alpha \), the resulting representation is called a **first-order (Taylor series) approximation** of \( f \) at (or about) the point \( \bar{x} \).

Note that if \( f \) is differentiable at \( \bar{x} \), there could only be one gradient vector, and this vector is given by

\[ \nabla f(\bar{x}) = \left( \frac{\partial f(\bar{x})}{\partial x_1}, ..., \frac{\partial f(\bar{x})}{\partial x_n} \right)' = (f_1(\bar{x}), ..., f_n(\bar{x}))', \]
where \( f_i(\bar{x}) = \frac{\partial f(\bar{x})}{\partial x_i} \) is the partial derivative of \( f \) with respect to \( x_i \) at \( \bar{x} \) (see Exercise 3.36, and review Appendix A.4).

The following lemma shows that a differentiable convex function has only one subgradient, the gradient vector. Hence, the results of the preceding section can easily be specialized to the differentiable case, in which the gradient vector replaces subgradients.

### 3.3.2 Lemma

Let \( S \) be a nonempty convex set in \( \mathbb{R}^n \), and let \( f: S \rightarrow \mathbb{R} \) be convex. Suppose that \( f \) is differentiable at \( \bar{x} \in \text{int} \ S \). Then the collection of subgradients of \( f \) at \( \bar{x} \) is the singleton set \( \{\nabla f(\bar{x})\} \).

**Proof**

By Theorem 3.2.5, the set of subgradients of \( f \) at \( \bar{x} \) is not empty. Now, let \( \xi \) be a subgradient of \( f \) at \( \bar{x} \). As a result of Theorem 3.2.5 and the differentiability of \( f \) at \( \bar{x} \), for any vector \( d \) and for \( \lambda \) sufficiently small, we get

\[
\begin{align*}
f(\bar{x} + \lambda d) & \geq f(\bar{x}) + \lambda \nabla f(\bar{x})^T d \\
f(\bar{x} + \lambda d) & = f(\bar{x}) + \lambda \nabla f(\bar{x})^T d + \lambda \|d\| \alpha(\bar{x}; \lambda d).
\end{align*}
\]

Subtracting the equation from the inequality, we obtain

\[
0 \geq \lambda [\xi - \nabla f(\bar{x})]^T d - \lambda \|d\| \alpha(\bar{x}; \lambda d).
\]

If we divide by \( \lambda > 0 \) and let \( \lambda \rightarrow 0^+ \), it follows that \( [\xi - \nabla f(\bar{x})]^T d \leq 0 \).

Choosing \( d = \xi - \nabla f(\bar{x}) \), the last inequality implies that \( \xi = \nabla f(\bar{x}) \). This completes the proof.

In the light of Lemma 3.3.2, we give the following important characterization of differentiable convex functions. The proof is immediate from Theorems 3.2.5 and 3.2.6 and Lemma 3.3.2.

### 3.3.3 Theorem

Let \( S \) be a nonempty open convex set in \( \mathbb{R}^n \), and let \( f: S \rightarrow \mathbb{R} \) be differentiable on \( S \). Then \( f \) is convex if and only if for any \( \bar{x} \in S \), we have

\[
f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) \quad \text{for each} \ x \in S.
\]

Similarly, \( f \) is strictly convex if and only if for each \( \bar{x} \in S \), we have

\[
f(x) > f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) \quad \text{for each} \ x \neq \bar{x} \text{ in} \ S.
\]

There are two evident implications of the above result that find use in various contexts. The first is that if we have an optimization problem to minimize \( f(x) \) subject to \( x \in X \), where \( f \) is a convex function, then given any point \( \bar{x} \), the affine
function \( f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) \) bounds \( f \) from below. Hence, the minimum of \( f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) \) over \( X \) (or over a relaxation of \( X \)) yields a lower bound on the optimum value of the given optimization problem, which can prove to be useful in an algorithmic approach. A second point in the same spirit is that this affine bounding function can be used to derive polyhedral outer approximations. For example, consider the set \( X = \{ x : g_i(x) \leq 0, \ i = 1, \ldots, m \} \), where \( g_i \) is a convex function for each \( i = 1, \ldots, m \). Given any point \( \bar{x} \), construct the polyhedral set \( \bar{X} = \{ x : g_i(\bar{x}) + \nabla g_i(\bar{x})^T (x - \bar{x}) \leq 0, \ i = 1, \ldots, m \} \). Note that the polyhedral set \( \bar{X} \) contains \( X \) and, hence, affords an outer linearization of this set, since for any \( x \in X \), we have \( 0 \geq g_i(x) \geq g_i(\bar{x}) + \nabla g_i(\bar{x})^T (x - \bar{x}) \) for \( i = 1, \ldots, m \) by Theorem 3.3.3. Such representations play a central role in many successive approximation algorithms for various nonlinear optimization problems.

The following theorem gives another necessary and sufficient characterization of differentiable convex functions. For a function of one variable, the characterization reduces to the slope being nondecreasing.

3.3.4 Theorem

Let \( S \) be a nonempty open convex set in \( \mathbb{R}^n \) and let \( f : S \to \mathbb{R} \) be differentiable on \( S \). Then \( f \) is convex if and only if for each \( x_1, x_2 \in S \) we have

\[
[\nabla f(x_2) - \nabla f(x_1)]^T (x_2 - x_1) \geq 0.
\]

Similarly, \( f \) is strictly convex if and only if, for each distinct \( x_1, x_2 \in S \), we have

\[
[\nabla f(x_2) - \nabla f(x_1)]^T (x_2 - x_1) > 0.
\]

Proof

Assume that \( f \) is convex, and let \( x_1, x_2 \in S \). By Theorem 3.3.3 we have

\[
f(x_1) \geq f(x_2) + \nabla f(x_2)^T (x_1 - x_2)
\]

\[
f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1).
\]

Adding the two inequalities, we get \([\nabla f(x_2) - \nabla f(x_1)]^T (x_2 - x_1) \geq 0\). To show the converse, let \( x_1, x_2 \in S \). By the mean value theorem,

\[
f(x_2) - f(x_1) = \nabla f(x)^T (x_2 - x_1),
\]

where \( x = \lambda x_1 + (1 - \lambda)x_2 \) for some \( \lambda \in (0, 1) \). By assumption, \([\nabla f(x) - \nabla f(x_1)]^T (x - x_1) \geq 0\); that is, \((1 - \lambda)[\nabla f(x) - \nabla f(x_1)]^T (x_2 - x_1) \geq 0\). This implies that
\( \nabla f(x)'(x_2 - x_1) \geq \nabla f(x_1)'(x_2 - x_1) \). By (3.11) we get \( f(x_2) \geq f(x_1) + \nabla f(x_1)'(x_2 - x_1) \), so by Theorem 3.3.3, \( f \) is convex. The strict case is similar and the proof is complete.

Even though Theorems 3.3.3 and 3.3.4 provide necessary and sufficient characterizations of convex functions, checking these conditions is difficult from a computational standpoint. A simple and more manageable characterization, at least for quadratic functions, can be obtained, provided that the function is twice differentiable.

**Twice Differentiable Convex and Concave Functions**

A function \( f \) that is differentiable at \( \bar{x} \) is said to be twice differentiable at \( \bar{x} \) if the second-order (Taylor series) expansion representation of Definition 3.3.5 exists.

**3.3.5 Definition**

Let \( S \) be a nonempty set in \( \mathbb{R}^n \), and let \( f: S \rightarrow \mathbb{R} \). Then \( f \) is said to be twice differentiable at \( x \in \text{int} \ S \) if there exist a vector \( \nabla f(\bar{x}) \), and an \( n \times n \) symmetric matrix \( H(\bar{x}) \), called the Hessian matrix, and a function \( \alpha: \mathbb{R}^n \rightarrow \mathbb{R} \) such that

\[
f(x) = f(\bar{x}) + \nabla f(\bar{x})'(x - \bar{x}) + \frac{1}{2}(x - \bar{x})'H(\bar{x})(x - \bar{x}) + \|x - \bar{x}\|^2 \alpha(\bar{x}; x - \bar{x})
\]

for each \( x \in S \), where \( \lim_{x \to \bar{x}} \alpha(\bar{x}; x - \bar{x}) = 0 \). The function \( f \) is said to be twice differentiable on the open set \( S' \subseteq S \) if it is twice differentiable at each point in \( S' \).

It may be noted that for twice differentiable functions, the Hessian matrix \( H(\bar{x}) \) is comprised of the second-order partial derivatives \( f_{ij}(\bar{x}) = \partial^2 f(\bar{x}) / \partial x_i \partial x_j \) for \( i = 1, \ldots, n, j = 1, \ldots, n \), and is given as follows:

\[
H(\bar{x}) = \begin{bmatrix}
f_{11}(\bar{x}) & f_{12}(\bar{x}) & \cdots & f_{1n}(\bar{x}) \\
f_{21}(\bar{x}) & f_{22}(\bar{x}) & \cdots & f_{2n}(\bar{x}) \\
\vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots \\
f_{n1}(\bar{x}) & f_{n2}(\bar{x}) & \cdots & f_{nn}(\bar{x})
\end{bmatrix}
\]

In expanded form, the foregoing representation can be written as
Convex Functions and Generalizations

\[ f(x) = f(\bar{x}) + \sum_{j=1}^{n} f_j(\bar{x})(x_j - \bar{x}_j) + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_j - \bar{x}_j)(x_j - \bar{x}_j) f_{ij}(\bar{x}) + \|x - \bar{x}\|^2 \alpha(\bar{x}; x - \bar{x}). \]

Again, without the remainder term associated with the function \( \alpha \), this representation is known as a second-order (Taylor series) approximation at (or about) the point \( \bar{x} \).

### 3.3.6 Examples

**Example 1.** Let \( f(x_1, x_2) = 2x_1 + 6x_2 - 2x_1^2 - 3x_2^2 + 4x_1x_2 \). Then we have

\[
\nabla f(\bar{x}) = \begin{bmatrix} 2 - 4\bar{x}_1 + 4\bar{x}_2 \\ 6 - 6\bar{x}_2 + 4\bar{x}_1 \end{bmatrix} \quad \text{and} \quad H(\bar{x}) = \begin{bmatrix} -4 & 4 \\ 4 & -6 \end{bmatrix}.
\]

For example, taking \( \bar{x} = (0, 0)' \), the second-order expansion of this function is given by

\[
f(x_1, x_2) = (2, 6) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2} (x_1, x_2) \begin{bmatrix} -4 & 4 \\ 4 & -6 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.
\]

Note that there is no remainder term here since the given function is quadratic, so the above representation is exact.

**Example 2.** Let \( f(x_1, x_2) = e^{2x_1 + 3x_2} \). Then we get

\[
\nabla f(\bar{x}) = \begin{bmatrix} 2e^{2\bar{x}_1 + 3\bar{x}_2} \\ 3e^{2\bar{x}_1 + 3\bar{x}_2} \end{bmatrix} \quad \text{and} \quad H(\bar{x}) = \begin{bmatrix} 4e^{2\bar{x}_1 + 3\bar{x}_2} & 6e^{2\bar{x}_1 + 3\bar{x}_2} \\ 6e^{2\bar{x}_1 + 3\bar{x}_2} & 9e^{2\bar{x}_1 + 3\bar{x}_2} \end{bmatrix}.
\]

Hence, the second-order expansion of this function about the point \( \bar{x} = (2, 1)' \) is given by

\[
f(\bar{x}) = e^7 + (2e^7, 3e^7) \begin{pmatrix} x_1 - 2 \\ x_2 - 1 \end{pmatrix} + \frac{1}{2} (x_1 - 2, x_2 - 1) \begin{bmatrix} 4e^7 & 6e^7 \\ 6e^7 & 9e^7 \end{bmatrix} \begin{pmatrix} x_1 - 2 \\ x_2 - 1 \end{pmatrix}
\]

\[+ \|x - \bar{x}\|^2 \alpha(\bar{x}; x - \bar{x}).\]

Theorem 3.3.7 shows that \( f \) is convex on \( S \) if and only if its Hessian matrix is positive semidefinite (PSD) everywhere in \( S \); that is, for any \( \bar{x} \) in \( S \), we have \( x'H(\bar{x})x \geq 0 \) for all \( x \in \mathbb{R}^n \). Symmetrically, a function \( f \) is concave on \( S \) if and only if its Hessian matrix is negative semidefinite (NSD) everywhere in \( S \),
that is, for any $\bar{x} \in S$, we have $x' H(\bar{x}) x \leq 0$ for all $x \in \mathbb{R}^n$. A matrix that is neither positive nor negative semidefinite is called *indefinite* (ID).

### 3.3.7 Theorem

Let $S$ be a nonempty open convex set in $\mathbb{R}^n$, and let $f : S \to \mathbb{R}$ be twice differentiable on $S$. Then $f$ is convex if and only if the Hessian matrix is positive semidefinite at each point in $S$.

#### Proof

Suppose that $f$ is convex, and let $\bar{x} \in S$. We need to show that $x' H(\bar{x}) x \geq 0$ for each $x \in \mathbb{R}^n$. Since $S$ is open, then for any given $x \in \mathbb{R}^n$, $\bar{x} + \lambda x \in S$ for $|\lambda| \neq 0$ and sufficiently small. By Theorem 3.3.3 and by the twice differentiability of $f$, we get the following two expressions:

$$f(\bar{x} + \lambda x) \geq f(\bar{x}) + \lambda \nabla f(\bar{x})' x$$

$$f(\bar{x} + \lambda x) = f(\bar{x}) + \lambda \nabla f(\bar{x})' x + \frac{1}{2} \lambda^2 x' H(\bar{x}) x + \lambda^2 \|x\|^2 \alpha(\bar{x}; \lambda x).$$

Subtracting (3.13) from (3.12), we get

$$\frac{1}{2} \lambda^2 x' H(\bar{x}) x + \lambda^2 \|x\|^2 \alpha(\bar{x}; \lambda x) \geq 0.$$ 

Dividing by $\lambda^2 > 0$ and letting $\lambda \to 0$, it follows that $x' H(\bar{x}) x \geq 0$. Conversely, suppose that the Hessian matrix is positive semidefinite at each point in $S$. Consider $x$ and $\bar{x}$ in $S$. Then, by the mean value theorem, we have

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})' (x - \bar{x}) + \frac{1}{2} (x - \bar{x})' H(\hat{x})(x - \bar{x}),$$

where $\hat{x} = \lambda \bar{x} + (1 - \lambda)x$ for some $\lambda \in (0, 1)$. Note that $\hat{x} \in S$ and hence, by assumption, $H(\hat{x})$ is positive semidefinite. Therefore, $(x - \bar{x})' H(\hat{x})(x - \bar{x}) \geq 0$, and from (3.14), we conclude that

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})' (x - \bar{x}).$$

Since the above inequality is true for each $x$, $\bar{x}$ in $S$, $f$ is convex by Theorem 3.3.3. This completes the proof.

Theorem 3.3.7 is useful in checking the convexity or concavity of a twice differentiable function. In particular, if the function is quadratic, the Hessian matrix is independent of the point under consideration. Hence, checking its convexity reduces to checking the positive semidefiniteness of a constant matrix.
Results analogous to Theorem 3.3.7 can be obtained for the strict convex and concave cases. It turns out that if the Hessian matrix is positive definite at each point in $S$, the function is strictly convex. In other words, if for any given point $\mathbf{x}$ in $S$, we have $\mathbf{x}^T H(\mathbf{x}) \mathbf{x} > 0$ for all $\mathbf{x} \neq 0$ in $\mathbb{R}^n$, then $f$ is strictly convex. This follows readily from the proof of Theorem 3.3.7. However, if $f$ is strictly convex, its Hessian matrix is positive semidefinite, but not necessarily positive definite everywhere in $S$, unless, for example, if $f$ is quadratic. The latter is seen by writing (3.12) as a strict inequality for $\mathbf{x} \neq 0$ and noting that the remainder term in (3.13) is then absent. To illustrate, consider the strictly convex function defined by $f(x) = x^4$. The Hessian matrix $H(x) = 12x^2$ is positive definite for all nonzero $x$ but is positive semidefinite, and not positive definite, at $x = 0$. The following theorem records this fact.

3.3.8 Theorem

Let $S$ be a nonempty open convex set in $\mathbb{R}^n$, and let $f: S \to \mathbb{R}$ be twice differentiable on $S$. If the Hessian matrix is positive definite at each point in $S$, $f$ is strictly convex. Conversely, if $f$ is strictly convex, the Hessian matrix is positive semidefinite at each point in $S$. However, if $f$ is strictly convex and quadratic, its Hessian is positive definite.

The foregoing result can be strengthened somewhat while providing some additional insights into the second-order characterization of convexity. Consider, for example, the univariate function $f(x) = x^4$ addressed above, and let us show how we can argue that this function is strictly convex despite the fact that $f''(0) = 0$. Since $f''(x) \geq 0$ for all $x \in \mathbb{R}$, we have by Theorem 3.3.7 that $f$ is convex. Hence, by Theorem 3.3.3, all that we need to show is that for any point $\mathbf{x}$, the supporting hyperplane $y = f(\mathbf{x}) + f'(\mathbf{x})(x - \mathbf{x})$ to the epigraph of the function touches this epigraph only at the given point $(x, y) = (\mathbf{x}, f(\mathbf{x}))$. On the contrary, if this supporting hyperplane also touches the epigraph at some other point $(\mathbf{\hat{x}}, f(\mathbf{\hat{x}}))$, we have $f(\mathbf{\hat{x}}) = f(\mathbf{x}) + f'(\mathbf{x})(\mathbf{\hat{x}} - \mathbf{x})$. But this means that for any $x = \lambda \mathbf{x} + (1 - \lambda)\mathbf{\hat{x}}$, $0 \leq \lambda \leq 1$, we have, upon using Theorem 3.3.3 and the convexity of $f$,

$$\lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{\hat{x}}) = f(\mathbf{x}) + f'(\mathbf{x})(x - \mathbf{x}) \leq f(x) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{\hat{x}}).$$

Hence, equality holds true throughout, and the supporting hyperplane touches the graph of the function at all convex combinations $(x, f(x))$ as well. In fact, we obtain $f(x) = \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{\hat{x}})$ for all $0 \leq \lambda \leq 1$, so $f''(x) = 0$ at the uncountably infinite number of points $x$ for all $0 < \lambda < 1$. This contradicts the fact that $f''(x) = 0$ only at $x = 0$ from the above example, and therefore, the function is strictly convex. As a result, if we lose positive definiteness of a...
univariate convex function at only a finite (or countably infinite) number of points, we can still claim that this function is strictly convex.

Staying with univariate functions for the time being, if the function is infinitely differentiable, we can derive a necessary and sufficient condition for the function to be strictly convex. [By an infinitely differentiable function \( f: \mathbb{R}^n \rightarrow \mathbb{R} \), we mean one for which for any \( \bar{x} \) in \( \mathbb{R}^n \), derivatives of all orders exist and so are continuous; are uniformly bounded in values; and for which the infinite Taylor series expansion of \( f(x) \) about \( f(\bar{x}) \) gives an infinite series representation of the value of \( f \). Of course, this infinite series can possibly have only a finite number of terms, as, for example, when derivatives of order exceeding some value all vanish.]

### 3.3.9 Theorem

Let \( S \) be a nonempty open convex set in \( \mathbb{R} \), and let \( f: S \rightarrow \mathbb{R} \) be infinitely differentiable. Then \( f \) is strictly convex on \( S \) if and only if for each \( \bar{x} \in S \), there exists an even \( n \) such that \( f^{(n)}(\bar{x}) > 0 \), while \( f^{(j)}(\bar{x}) = 0 \) for any \( 1 < j < n \), where \( f^{(j)} \) denotes the \( j \)th-order derivative of \( f \).

**Proof**

Let \( \bar{x} \) be any point in \( S \), and consider the infinite Taylor series expansion of \( f \) about \( \bar{x} \) for a perturbation \( h \neq 0 \) and small enough:

\[
f(\bar{x} + h) = f(\bar{x}) + hf'(\bar{x}) + \frac{h^2}{2!} f''(\bar{x}) + \frac{h^3}{3!} f'''(\bar{x}) + \cdots
\]

If \( f \) is strictly convex, then by Theorem 3.3.3 we have that \( f(\bar{x} + h) > f(\bar{x}) + hf'(\bar{x}) \) for \( h \neq 0 \). Using this above, we get that for all \( h \neq 0 \) and sufficiently small,

\[
\frac{h^2}{2!} f''(\bar{x}) + \frac{h^3}{3!} f'''(\bar{x}) + \frac{h^4}{4!} f''''(\bar{x}) + \cdots > 0.
\]

Hence, not all derivatives of order greater than or equal to 2 at \( \bar{x} \) can be zero. Moreover, since by making \( h \) sufficiently small, we can make the first nonzero term above dominate the rest of the expansion, and since \( h \) can be of either sign, it follows that this first nonzero derivative must be of an even order and positive for the inequality to hold true.

Conversely, suppose that given any \( \bar{x} \in S \), there exists an even \( n \) such that \( f^{(n)}(\bar{x}) > 0 \), while \( f^{(j)}(\bar{x}) = 0 \) for \( 1 < j < n \). Then, as above, we have \( (\bar{x} + h) \in S \) and \( f(\bar{x} + h) > f(\bar{x}) + hf'(\bar{x}) \) for all \( -\delta < h < \delta \), for some \( \delta > 0 \) and sufficiently small. Now the hypothesis given also asserts that \( f''(\bar{x}) \geq 0 \) for all \( \bar{x} \in S \), so by Theorem 3.3.7 we know that \( f \) is convex. Consequently, for any \( h \neq 0 \), with \( (\bar{x} + \bar{h}) \in S \), we get \( f(\bar{x} + \bar{h}) \geq f(\bar{x}) + \bar{h}f'(\bar{x}) \) by Theorem 3.3.3. To
complete the proof, we must show that this inequality is indeed strict. On the contrary, if \( f(x + h) = f(x) + hf'(x) \), we get

\[
\lambda f(x + h) + (1 - \lambda)f(x) = f(x) + \lambda hf'(x) \leq f(x + \lambda h) = f[\lambda(x + h) + (1 - \lambda)x] \leq \lambda f(x + h) + (1 - \lambda)f(x)
\]

for all \( 0 \leq \lambda \leq 1 \). But this means that equality holds throughout and that \( f(x + \lambda h) = f(x) + \lambda hf'(x) \) for all \( 0 \leq \lambda \leq 1 \). By taking \( \lambda \) close enough to zero, we can contradict the statement that \( f(x + h) > f(x) + hf'(x) \) for all \( -\delta < h < \delta \), and this completes the proof.

To illustrate, when \( f(x) = x^4 \), we have \( f'(x) = 4x^3 \) and \( f''(x) = 12x^2 \). Hence, for \( x \neq 0 \), the first nonzero derivative as in Theorem 3.3.9 is of order 2 and is positive. Furthermore, for \( x = 0 \), we have \( f''(x) = f''(0) = 0 \) and \( f^{(4)}(x) = 24 > 0 \); so by Theorem 3.3.9, we can conclude that \( f \) is strictly convex.

Now let us turn to the multivariate case. The following result provides an insightful connection between the univariate and multivariate cases and permits us to derive results for the latter case from those for the former case. For notational simplicity, we have stated this result for \( \mathbb{R}^n \rightarrow \mathbb{R} \), although one can readily restate it for \( \mathbb{S} \rightarrow \mathbb{R} \), where \( \mathbb{S} \) is some nonempty convex subset of \( \mathbb{R}^n \).

3.3.10 Theorem

Consider a function \( f: \mathbb{R}^n \rightarrow \mathbb{R} \), and for any point \( x \in \mathbb{R}^n \) and a nonzero direction \( d \in \mathbb{R}^n \), define \( F(\lambda) = f(\lambda x + \lambda d) \) as a function of \( \lambda \in \mathbb{R} \). Then \( f \) is (strictly) convex if and only if \( F(\lambda) \) is (strictly) convex for all \( x \) and \( d \neq 0 \) in \( \mathbb{R}^n \).

Proof

Given any \( x \) and \( d \neq 0 \) in \( \mathbb{R}^n \), let us write \( F(\lambda) \) simply as \( F(\lambda) \) for convenience. If \( f \) is convex, then for any \( \lambda_1 \) and \( \lambda_2 \) in \( \mathbb{R} \) and for any \( 0 \leq \alpha \leq 1 \), we have

\[
F(\alpha \lambda_1 + (1 - \alpha) \lambda_2) = f(\alpha \lambda_1 d + (1 - \alpha) \lambda_2 d) \leq \alpha F(\lambda_1) + (1 - \alpha)F(\lambda_2).
\]

Hence, \( F \) is convex. Conversely, suppose that \( F(\lambda), \lambda \in \mathbb{R} \), is convex for all \( x \) and \( d \neq 0 \) in \( \mathbb{R}^n \). Then, for any \( x_1 \) and \( x_2 \) in \( \mathbb{R}^n \) and \( 0 \leq \lambda \leq 1 \), we have
\[ \lambda f(x_1) + (1 - \lambda) f(x_2) = \lambda f(x_1 + 0(x_2 - x_1)) + (1 - \lambda) f(x_1 + 1(x_2 - x_1)) \]
\[ = \lambda f_{x_1}(x_2 - x_1)(0) + (1 - \lambda) f_{x_1}(x_2 - x_1)(1) \]
\[ \geq f_{x_1}(x_2 - x_1)(1 - \lambda) \]
\[ = f(x_1 + (1 - \lambda)(x_2 - x_1)) = f(\lambda x_1 + (1 - \lambda)x_2), \]

so \( f \) is convex. The argument for the strictly convex case is similar, and this completes the proof.

This insight of examining \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) via its univariate cross sections \( F(\bar{x},d) \) can be very useful both as a conceptual tool for viewing \( f \) and as an analytical tool for deriving various results. For example, writing \( F(\lambda) = F(\bar{x},d)(\lambda) = f(\bar{x} + \lambda d) \), for any given \( \bar{x} \) and \( d \neq 0 \) in \( \mathbb{R}^n \), we have from the univariate Taylor series expansion (assuming infinite differentiability) that

\[ F(\lambda) = F(0) + \lambda F'(0) + \frac{\lambda^2}{2!} F''(0) + \frac{\lambda^3}{3!} F'''(0) + \cdots. \]

By using the chain rule for differentiation, we obtain

\[ F'(\lambda) = \nabla f(\bar{x} + \lambda d)^t d = \sum_i f_i(\bar{x} + \lambda d) d_i \]
\[ F''(\lambda) = d^t H(\bar{x} + \lambda d) d = \sum_{i,j} f_{ij}(\bar{x} + \lambda d) d_i d_j \]
\[ F'''(\lambda) = \sum_{i,j,k} f_{ijk}(\bar{x} + \lambda d) d_i d_j d_k, \text{ etc.} \]

Substituting above, this gives the corresponding multivariate Taylor series expansion as

\[ f(\bar{x} + \lambda d) = f(\bar{x}) + \lambda \nabla f(\bar{x})^t d + \frac{\lambda^2}{2!} d^t H(\bar{x}) d + \frac{\lambda^3}{3!} \sum_{i,j,k} f_{ijk}(\bar{x}) d_i d_j d_k + \cdots. \]

As another example, using the second-order derivative result for characterizing the convexity of a univariate function along with Theorem 3.3.10, we can derive that \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) is convex if and only if \( F''_{\bar{x},d}(\lambda) \geq 0 \) for all \( \lambda \in \mathbb{R}, \bar{x} \in \mathbb{R}^n \), and \( d \in \mathbb{R}^n \). But since \( \bar{x} \) and \( d \) can be chosen arbitrarily, this is equivalent to requiring that \( F''_{\bar{x},d}(0) \geq 0 \) for all \( \bar{x} \) and \( d \) in \( \mathbb{R}^n \). From above, this translates to the statement that \( d^t H(\bar{x}) d \geq 0 \) for all \( d \in \mathbb{R}^n \), for each \( \bar{x} \in \mathbb{R}^n \), or that \( H(\bar{x}) \) is positive semidefinite for all \( \bar{x} \in \mathbb{R}^n \), as in Theorem 3.3.7. In a similar manner, or by using the multivariate Taylor series expansion directly as in the proof of
Theorem 3.3.9, we can assert that an infinitely differentiable function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is strictly convex if and only if for each \( \mathbf{x} \) and \( \mathbf{d} \neq \mathbf{0} \) in \( \mathbb{R}^n \), the first nonzero derivative term \( [F^{(j)}(0)] \) of order greater than or equal to 2 in the Taylor series expansion above exists, is of even order, and is positive. We leave the details of exploring this result to the reader in Exercise 3.38.

We present below an efficient (polynomial-time) algorithm for checking the definiteness of a (symmetric) Hessian matrix \( \mathbf{H}(\mathbf{x}) \) using elementary Gauss–Jordan operations. Appendix A cites a characterization of definiteness in terms of eigenvalues which finds use in some analytical proofs but is not an algorithmically convenient alternative. Moreover, if one needs to check for the definiteness of a matrix \( \mathbf{H}(\mathbf{x}) \) that is a function of \( \mathbf{x} \), this eigenvalue method is very cumbersome, if not virtually impossible, to use. Although the method presented below can also get messy in such instances, it is overall a more simple and efficient approach.

We begin by considering a \( 2 \times 2 \) Hessian matrix \( \mathbf{H} \) in Lemma 3.3.11, where the argument \( \mathbf{x} \) has been suppressed for convenience. This is then generalized in an inductive fashion to an \( n \times n \) matrix in Theorem 3.3.12.

3.3.11 Lemma

Consider a symmetric matrix \( \mathbf{H} = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \). Then \( \mathbf{H} \) is positive semidefinite if and only if \( a \geq 0 \), \( c \geq 0 \), and \( ac - b^2 \geq 0 \), and is positive definite if and only if the foregoing inequalities are all strict.

**Proof**

By definition, \( \mathbf{H} \) is positive semidefinite if and only if \( \mathbf{d}' \mathbf{H} \mathbf{d} = ad_1^2 + 2bd_1d_2 + cd_2^2 \geq 0 \) for all \( (d_1, d_2)' \in \mathbb{R}^2 \). Hence, if \( \mathbf{H} \) is positive semidefinite, we must clearly have \( a \geq 0 \) and \( c \geq 0 \). Moreover, if \( a = 0 \), we must have \( b = 0 \), so \( ac - b^2 = 0 \); or else, by taking \( d_2 = 1 \) and \( d_1 = -Mb \) for \( M > 0 \) and large enough, we would obtain \( \mathbf{d}' \mathbf{H} \mathbf{d} < 0 \), a contradiction. On the other hand, if \( a > 0 \), then completing the squares, we get

\[
\mathbf{d}' \mathbf{H} \mathbf{d} = a \left( d_1^2 + \frac{2bd_1d_2}{a} + \frac{b^2}{a^2} d_2^2 \right) + d_2^2 \left( c - \frac{b^2}{a} \right) = a \left( d_1 + \frac{b}{a} d_2 \right)^2 + d_2^2 \left( \frac{ac - b^2}{a} \right).
\]

Hence, we must again have \( (ac - b^2) \geq 0 \), since otherwise, by taking \( d_2 = 1 \) and \( d_1 = -b/a \), we would get \( \mathbf{d}' \mathbf{H} \mathbf{d} = (ac - b^2)/a < 0 \), a contradiction. Hence, the condition of the theorem holds true. Conversely, suppose that \( a \geq 0 \), \( c \geq 0 \), and
\( ac - b^2 \geq 0 \). If \( a = 0 \), this gives \( b = 0 \), so \( \mathbf{d}'\mathbf{H}\mathbf{d} = cd_2^2 \geq 0 \). On the other hand, if \( a > 0 \), by completing the squares as above we get

\[
\mathbf{d}'\mathbf{H}\mathbf{d} = a \left( d_1 + \frac{b}{a} d_2^2 \right) + d_2^2 \left( \frac{ac - b^2}{a} \right) \geq 0.
\]

Hence, \( \mathbf{H} \) is positive semidefinite. The proof of positive definiteness is similar, and this completes the proof.

We remark here that since a matrix \( \mathbf{H} \) is negative semidefinite (negative definite) if and only if \(-\mathbf{H}\) is positive semidefinite (positive definite), we get from Lemma 3.3.11 that \( \mathbf{H} \) is negative semidefinite if and only if \( a \leq 0 \), \( c \leq 0 \), and \( ac - b^2 \geq 0 \), and that \( \mathbf{H} \) is negative definite if and only if these inequalities are all strict. Theorem 3.3.12 is stated for checking positive semidefiniteness or positive definiteness of \( \mathbf{H} \). By replacing \( \mathbf{H} \) by \(-\mathbf{H}\), we could test symmetrically for negative semidefiniteness or negative definiteness. If the matrix turns out to be neither positive semidefinite nor negative semidefinite, it is indefinite. Also, we assume below that \( \mathbf{H} \) is symmetric, being the Hessian of a twice differentiable function for our purposes. In general, if \( \mathbf{H} \) is not symmetric, then since \( \mathbf{d}'\mathbf{H}\mathbf{d} = \mathbf{d}'\mathbf{H}'\mathbf{d} = \mathbf{d}'\left( \frac{(\mathbf{H} + \mathbf{H}')}{2} \right)\mathbf{d} \), we can check for the definiteness of \( \mathbf{H} \) by using the symmetric matrix \( \left( \frac{\mathbf{H} + \mathbf{H}'}{2} \right) \) below.

**3.3.12 Theorem (Checking for PSD/PD)**

Let \( \mathbf{H} \) be a symmetric \( n \times n \) matrix with elements \( h_{ij} \).

(a) If \( h_{ii} \leq 0 \) for any \( i \in \{1, \ldots, n\} \), \( \mathbf{H} \) is not positive definite; and if \( h_{ii} < 0 \) for any \( i \in \{1, \ldots, n\} \), \( \mathbf{H} \) is not positive semidefinite.

(b) If \( h_{ii} = 0 \) for any \( i \in \{1, \ldots, n\} \), we must have \( h_{ij} = h_{ji} = 0 \) for all \( j = 1, \ldots, n \) as well, or else \( \mathbf{H} \) is not positive semidefinite.

(c) If \( n = 1 \), \( \mathbf{H} \) is positive semidefinite (positive definite) if and only if \( h_{11} \geq 0 \) (> 0). Otherwise, if \( n \geq 2 \), let

\[
\mathbf{H} = \begin{bmatrix} h_{11} & \mathbf{q}' \\ \mathbf{q} & \mathbf{G} \end{bmatrix}
\]

in partitioned form, where \( \mathbf{q} = 0 \) if \( h_{11} = 0 \), and otherwise, \( h_{11} > 0 \). Perform elementary Gauss–Jordan operations using the first row of \( \mathbf{H} \) to reduce it to the following matrix in either case:

\[
\mathbf{H} = \begin{bmatrix} h_{11} & \mathbf{q}' \\ 0 & \mathbf{G}_{\text{new}} \end{bmatrix}
\]
Then $G_{\text{new}}$ is a symmetric $(n-1) \times (n-1)$ matrix, and $H$ is positive semidefinite if and only if $G_{\text{new}}$ is positive semidefinite. Moreover, if $h_1 > 0$, $H$ is positive definite if and only if $G_{\text{new}}$ is positive definite.

**Proof**

(a) Since $d' \! H \! d = d_i^2 h_{ii}$ whenever $d_j = 0$ for all $j \neq i$, Part (a) of the theorem is obviously true.

(b) Suppose that for some $i \neq j$, we have $h_{ii} = 0$ and $h_{ij} \neq 0$. Then, by taking $d_k = 0$ for all $k \neq i$ or $j$, we get $d' \! H \! d = 2 h_{ij} d_i d_j + d_j^2 h_{jj}$, which can be made negative as in the proof of Lemma 3.3.11 by taking $d_j = 1$ and $d_i = -h_{ij} M$ for $M > 0$ and sufficiently large. This establishes Part (b).

(c) Finally, suppose that $H$ is given in partitioned form as in Part (c). If $n = 1$, the result is trivial. Otherwise, for $n \geq 2$, let $d' = (d_1, \delta')$. If $h_{11} = 0$, by assumption we also have $q = 0$, and then $G_{\text{new}} = G$. Moreover, in this case, $d' \! H \! d = \delta' G_{\text{new}} \delta$, so $H$ is positive semidefinite if and only if $G_{\text{new}}$ is positive semidefinite. On the other hand, if $h_{11} > 0$, we get

$$d' \! H \! d = (d_1, \delta') \begin{bmatrix} h_{11} & q' \\ q & G \end{bmatrix} \begin{pmatrix} d_1 \\ \delta \end{pmatrix} = d_1^2 h_{11} + 2 d_1 (q' \delta) + \delta' G \delta.$$ 

But by the Gauss–Jordan reduction process, we have

$$G_{\text{new}} = G - \frac{1}{h_{11}} \begin{bmatrix} q_1 q' \\ q_2 q' \\ \vdots \\ q_n q' \end{bmatrix} = G - \frac{1}{h_{11}} qq',$$

which is a symmetric matrix. By substituting this above, we get

$$d' \! H \! d = d_1^2 h_{11} + 2 d_1 (q' \delta) + \delta' \left( G_{\text{new}} + \frac{1}{h_{11}} qq' \right) \delta = \delta' G_{\text{new}} \delta + h_{11} \left( d_1 + \frac{q' \delta}{h_{11}} \right)^2.$$

Hence, it can readily be verified that $d' \! H \! d \geq 0$ for all $d \in \mathbb{R}^n$ if and only if $\delta' G_{\text{new}} \delta \geq 0$ for all $\delta \in \mathbb{R}^{n-1}$, because $h_{11} (d_1 + q' \delta / h_{11})^2 \geq 0$, and the latter term can be made zero by selecting $d_1 = -q' \delta / h_{11}$, if necessary. By the same argu-
ment, \( d'HD > 0 \) for all \( d \neq 0 \) in \( R^n \) if and only if \( \delta' G_{\text{new}} \delta > 0 \) for all \( \delta \neq 0 \) in \( R^{n-1} \), and this completes the proof.

Observe that Theorem 3.3.12 prompts a polynomial-time algorithm for checking the PSD/PD of a symmetric \( n \times n \) matrix \( H \). We first scan the diagonal elements to see if either condition (a) or (b) leads to the conclusion that the matrix is not PSD/PD. If this does not terminate the process, we perform the Gauss–Jordan reduction as in Part (c) and arrive at a matrix \( G_{\text{new}} \) of one lesser dimension for which we may now perform the same test as on \( H \). When \( G_{\text{new}} \) is finally a \( 2 \times 2 \) matrix, we can use Lemma 3.3.11, or we can continue to reduce it to a \( 1 \times 1 \) matrix and hence determine the PSD/PD of \( H \). Since each pass through the inductive step of the algorithm is of complexity \( O(n^2) \) (read as "of order \( n^2 \)") and meaning that the number of elementary arithmetic operations, comparison, etc., involved are bounded above by \( Kn^2 \) for some constant \( K \) and the number of inductive steps is of \( O(n) \), the overall process is of polynomial complexity \( O(n^3) \). Because the algorithm basically works toward reducing the matrix to an upper triangular matrix, it is sometimes called a superdiagonalization algorithm. This algorithm affords a proof for the following useful result, which can alternatively be proved using the eigenvalue characterization of definiteness (see Exercise 3.42).

**Corollary**

Let \( H \) be an \( n \times n \) symmetric matrix. Then \( H \) is positive definite if and only if it is positive semidefinite and nonsingular.

**Proof**

If \( H \) is positive definite, it is positive semidefinite; and since the superdiagonalization algorithm reduces the matrix \( H \) to an upper triangular matrix with positive diagonal elements via elementary row operations, \( H \) is nonsingular. Conversely, if \( H \) is positive semidefinite and nonsingular, the superdiagonalization algorithm must always encounter nonzero elements along the diagonal because \( H \) is nonsingular, and these must be positive because \( H \) is positive semidefinite. Hence, \( H \) is positive definite.

**3.3.13 Examples**

**Example 1.** Consider Example 1 of Section 3.3.6. Here we have

\[
H(x) = \begin{bmatrix} -4 & 4 \\ 4 & -6 \end{bmatrix}
\]
\[-H(x) = \begin{bmatrix} 4 & -4 \\ -4 & 6 \end{bmatrix}.\]

By Lemma 3.3.11 we conclude that \(-H(x)\) is positive definite, so \(H(x)\) is negative definite and the function \(f\) is strictly concave.

**Example 2.** Consider the function \(f(x_1, x_2) = x_1^3 + 2x_2^2\). Here we have

\[
\nabla f(x) = \begin{bmatrix} 3x_1^2 \\ 4x_2 \end{bmatrix} \quad \text{and} \quad H(x) = \begin{bmatrix} 6x_1 & 0 \\ 0 & 4 \end{bmatrix}.
\]

By Lemma 3.3.11, whenever \(x_1 < 0\), \(H(x)\) is indefinite. However, \(H(x)\) is positive definite for \(x_1 > 0\), so \(f\) is strictly convex over \(\{x : x_1 > 0\}\).

**Example 3.** Consider the matrix

\[
H = \begin{bmatrix} 2 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix}.
\]

Note that the matrix is not negative semidefinite. To check PSD/PD, apply the superdiagonalization algorithm and reduce \(H\) to

\[
\begin{bmatrix} 2 & 1 & 2 \\ 0 & 3/2 & 2 \\ 0 & 2 & 2 \end{bmatrix}
\]

which gives \(G_{\text{new}} = \begin{bmatrix} 3/2 & 2 \\ 2 & 2 \end{bmatrix}\).

Now the diagonals of \(G_{\text{new}}\) are positive, but \(\det(G_{\text{new}}) = -1\). Hence, \(H\) is not positive semidefinite. Alternatively, we could have verified this by continuing to reduce \(G_{\text{new}}\) to obtain the matrix

\[
\begin{bmatrix} 3/2 & 2 \\ 0 & -2/3 \end{bmatrix}.
\]

Since the resulting second diagonal element (i.e., the reduced \(G_{\text{new}}\)) is negative, \(H\) is not positive semidefinite. Since \(H\) is not negative semidefinite either, it is indefinite.

### 3.4 Minima and Maxima of Convex Functions

In this section we consider the problems of minimizing and maximizing a convex function over a convex set and develop necessary and/or sufficient conditions for optimality.
Minimizing a Convex Function

The case of maximizing a concave function is similar to that of minimizing a convex function. We develop the latter in detail and ask the reader to draw the analogous results for the concave case.

3.4.1 Definition

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and consider the problem to minimize $f(x)$ subject to $x \in S$. A point $x \in S$ is called a feasible solution to the problem. If $\bar{x} \in S$ and $f(x) \geq f(\bar{x})$ for each $x \in S$, $\bar{x}$ is called an optimal solution, a global optimal solution, or simply a solution to the problem. The collection of optimal solutions are called alternative optimal solutions. If $\bar{x} \in S$ and if there exists an $\varepsilon$-neighborhood $N_\varepsilon(\bar{x})$ around $\bar{x}$ such that $f(x) \geq f(\bar{x})$ for each $x \in S \cap N_\varepsilon(\bar{x})$, $\bar{x}$ is called a local optimal solution. Similarly, if $\bar{x} \in S$ and if $f(x) > f(\bar{x})$ for all $x \in S \cap N_\varepsilon(\bar{x})$, $x \neq \bar{x}$, for some $\varepsilon > 0$, $\bar{x}$ is called a strict local optimal solution. On the other hand, if $\bar{x} \in S$ is the only local minimum in $S \cap N_\varepsilon(\bar{x})$, for some $\varepsilon$-neighborhood $N_\varepsilon(\bar{x})$ around $\bar{x}$, $\bar{x}$ is called a strong or isolated local optimal solution. All these types of local optima or minima are sometimes also referred to as relative minima. Figure 3.6 illustrates instances of local and global minima for the problem of minimizing $f(x)$ subject to $x \in S$, where $f$ and $S$ are shown in the figure.

The points in $S$ corresponding to A, B, and C are also both strict and strong local minima, whereas those corresponding to the flat segment of the graph between D and E are local minima that are neither strict nor strong. Note that if $\bar{x}$ is a strong or isolated local minimum, it is also a strict minimum. To see this, consider the $\varepsilon$-neighborhood $N_\varepsilon(\bar{x})$ characterizing the strong local minimum nature of $\bar{x}$. Then we must also have $f(x) > f(\bar{x})$ for all $x \in S \cap$
because otherwise, suppose that there exists an \( \tilde{x} \in S \cap N_{\varepsilon}^c(\bar{x}) \) such that 
\[ f(\tilde{x}) = f(\bar{x}). \]
Note that \( \tilde{x} \) is an alternative optimal solution within \( S \cap N_{\varepsilon}^c(\bar{x}) \),
so there exists some \( 0 < \varepsilon' < \varepsilon \) such that 
\[ f(x) \geq f(\tilde{x}) \]
for all \( x \in S \cap N_{\varepsilon'}^c(\tilde{x}) \).
But this contradicts the isolated local minimum status of \( \bar{x} \), and hence \( \bar{x} \) must
also be a strict local minimum. On the other hand, a strict local minimum need
not be an isolated local minimum. Figure 3.7 illustrates two such instances. In
Figure 3.7a, \( S = \mathbb{R} \) and \( f(x) = 1 \) for \( x = 1 \) and is equal to 2 otherwise. Note that
the point of discontinuity \( \bar{x} = 1 \) of \( f \) is a strict local minimum but is not isolated,
since any \( \varepsilon \)-neighborhood about \( \bar{x} \) contains points other than \( \bar{x} = 1 \), all of which
are also local minima. Figure 3.7b illustrates another case in which \( f(x) = x^2 \), a
strictly convex function; but \( S = \{1/2^k, k = 0,1,2,\ldots\} \cup \{0\} \) is a nonconvex set.

Here, the point \( \bar{x} = 1/2^k \) for any integer \( k \geq 0 \) is an isolated and therefore a strict
local minimum because it can be captured as the unique feasible solution in
\( S \cap N_{\varepsilon}^c(\bar{x}) \) for some sufficiently small \( \varepsilon > 0 \). However, although \( \bar{x} = 0 \) is clearly
a strict local minimum (it is, in fact, the unique global minimum), it is not
isolated because any \( \varepsilon \)-neighborhood about \( \bar{x} = 0 \) contains other local minima
of the foregoing type.

Nonetheless, for optimization problems, \( \min\{f(x) : x \in S\} \), where \( f \) is a
convex function and \( S \) is a convex set, which are known as convex programming
problems and that are of interest to us in this section, a strict local minimum is
also a strong local minimum, as shown in Theorem 3.4.2 (see Exercise 3.47 for a
weaker sufficient condition). The principal result here is that each local
minimum of a convex program is also a global minimum. This fact is quite
useful in the optimization process, since it enables us to stop with a global
optimal solution if the search in the vicinity of a feasible point does not lead to
an improving feasible solution.

### 3.4.2 Theorem

Let \( S \) be a nonempty convex set in \( \mathbb{R}^n \), and let \( f : S \to \mathbb{R} \) be convex on \( S \).
Consider the problem to minimize \( f(x) \) subject to \( x \in S \). Suppose that \( \bar{x} \in S \) is
a local optimal solution to the problem.

---

---

Figure 3.7 Strict local minima are not necessarily strong local minima.
1. Then $\bar{x}$ is a global optimal solution.

2. If either $\bar{x}$ is a strict local minimum or $f$ is strictly convex, $\bar{x}$ is the unique global optimal solution and is also a strong local minimum.

**Proof**

Since $\bar{x}$ is a local optimal solution, there exists an $\varepsilon$-neighborhood $N_\varepsilon(\bar{x})$ around $\bar{x}$ such that

$$f(x) \geq f(\bar{x}) \quad \text{for each } x \in S \cap N_\varepsilon(\bar{x}). \quad (3.15)$$

By contradiction, suppose that $\bar{x}$ is not a global optimal solution so that $f(\bar{x}) < f(\hat{x})$ for some $\hat{x} \in S$. By the convexity of $f$, the following is true for each $0 \leq \lambda \leq 1$:

$$f(\lambda \hat{x} + (1 - \lambda)\bar{x}) \leq \lambda f(\hat{x}) + (1 - \lambda)f(\bar{x}) \leq f(\bar{x}) + (1 - \lambda)f(\bar{x}) = f(\bar{x}).$$

But for $\lambda > 0$ and sufficiently small, $\lambda \hat{x} + (1 - \lambda)\bar{x} \in S \cap N_\varepsilon(\bar{x})$. Hence, the above inequality contradicts (3.15), and Part 1 is proved.

Next, suppose that $\bar{x}$ is a strict local minimum. By Part 1 it is a global minimum. Now, on the contrary, if there exists an $\hat{x} \in S$ such that $f(\hat{x}) = f(\bar{x})$, then defining $x_\lambda = \lambda \hat{x} + (1 - \lambda)\bar{x}$ for $0 \leq \lambda \leq 1$, we have, by the convexity of $f$ and $S$ that $f(x_\lambda) \leq \lambda f(\hat{x}) + (1 - \lambda)f(\bar{x}) = f(\bar{x})$, and $x_\lambda \in S$ for all $0 \leq \lambda \leq 1$. By taking $\lambda \to 0^+$, since we can make $x_\lambda \in N_\varepsilon(\bar{x}) \cap S$ for any $\varepsilon > 0$, this contradicts the strict local optimality of $\bar{x}$. Hence, $\bar{x}$ is the unique global minimum. Therefore, it must also be an isolated local minimum, since any other local minimum in $N_\varepsilon(\bar{x}) \cap S$ for any $\varepsilon > 0$ would also be a global minimum, which is a contradiction.

Finally, suppose that $\bar{x}$ is a local optimal solution and that $f$ is strictly convex. Since strict convexity implies convexity, then by Part 1, $\bar{x}$ is a global optimal solution. By contradiction, suppose that $\bar{x}$ is not the unique global optimal solution, so that there exists an $x \in S, x \neq \bar{x}$ such that $f(x) = f(\bar{x})$. By strict convexity,

$$f\left(\frac{1}{2}x + \frac{1}{2}\bar{x}\right) < \frac{1}{2}f(x) + \frac{1}{2}f(\bar{x}) = f(\bar{x}).$$

By the convexity of $S$, $(1/2)x + (1/2)\bar{x} \in S$, and the above inequality violates the global optimality of $\bar{x}$. Hence, $\bar{x}$ is the unique global minimum and, as above, is also a strong local minimum. This completes the proof.

We now develop a necessary and sufficient condition for the existence of a global solution. If such an optimal solution does not exist, then $\inf\{f(x) : x \in S\}$ is finite but is not achieved at any point in $S$, or it is equal to $-\infty$. 
3.4.3 Theorem

Let \( f: \mathbb{R}^n \to \mathbb{R} \) be a convex function, and let \( S \) be a nonempty convex set in \( \mathbb{R}^n \). Consider the problem to minimize \( f(x) \) subject to \( x \in S \). The point \( \bar{x} \in S \) is an optimal solution to this problem if and only if \( f \) has a subgradient \( \xi \) at \( \bar{x} \) such that \( \xi^t(x - \bar{x}) \geq 0 \) for all \( x \in S \).

Proof

Suppose that \( \xi^t(x - \bar{x}) \geq 0 \) for all \( x \in S \), where \( \xi \) is a subgradient of \( f \) at \( \bar{x} \). By the convexity of \( f \), we have
\[
f(x) \geq f(\bar{x}) + \xi^t(x - \bar{x}) \geq f(\bar{x})
\]
for all \( x \in S \), and hence \( \bar{x} \) is an optimal solution to the given problem.

To show the converse, suppose that \( \bar{x} \) is an optimal solution to the problem and construct the following two sets in \( \mathbb{R}^{n+1} \):
\[
\Lambda_1 = \{(x - \bar{x}, y) : x \in \mathbb{R}^n, y > f(x) - f(\bar{x})\}
\]
\[
\Lambda_2 = \{(x - \bar{x}, y) : x \in S, y \leq 0\}.
\]
The reader may easily verify that both \( \Lambda_1 \) and \( \Lambda_2 \) are convex sets. Also, \( \Lambda_1 \cap \Lambda_2 = \emptyset \) because otherwise there would exist a point \((x, y)\) such that
\[
x \in S, \quad 0 \geq y > f(x) - f(\bar{x}),
\]
contradicting the assumption that \( \bar{x} \) is an optimal solution to the problem. By Theorem 2.4.8 there is a hyperplane that separates \( \Lambda_1 \) and \( \Lambda_2 \); that is, there exist a nonzero vector \((\xi_0, \mu)\) and a scalar \( \alpha \) such that
\[
\xi_0^t(x - \bar{x}) + \mu y \leq \alpha, \quad \forall x \in \mathbb{R}^n, \quad y > f(x) - f(\bar{x}) \tag{3.16}
\]
\[
\xi_0^t(x - \bar{x}) + \mu y \geq \alpha, \quad \forall x \in S, \quad y \leq 0. \tag{3.17}
\]
If we let \( x = \bar{x} \) and \( y = 0 \) in (3.17), it follows that \( \alpha \leq 0 \). Next, letting \( x = \bar{x} \) and \( y = \varepsilon > 0 \) in (3.16), it follows that \( \mu \varepsilon \leq \alpha \). Since this is true for every \( \varepsilon > 0 \), \( \mu \leq 0 \) and \( \alpha \geq 0 \). To summarize, we have shown that \( \mu \leq 0 \) and \( \alpha = 0 \). If \( \mu = 0 \), from (3.16), \( \xi_0^t(x - \bar{x}) \leq 0 \) for each \( x \in \mathbb{R}^n \). If we let \( x = \bar{x} + \xi_0 \), it follows that
\[
0 \geq \xi_0^t(x - \bar{x}) = \|\xi_0\|^2
\]
and hence \( \xi_0 = 0 \). Since \((\xi_0, \mu) \neq (0, 0)\), we must have \( \mu < 0 \). Dividing (3.16) and (3.17) by \(-\mu\) and denoting \(-\xi_0/\mu\) by \( \xi \), we get the following inequalities:
\[ y \geq \xi^t(x - \overline{x}), \quad \forall x \in \mathbb{R}^n, \quad y > f(x) - f(\overline{x}) \quad (3.18) \]

\[ \xi^t(x - \overline{x}) - y \geq 0, \quad \forall x \in S, \quad y \leq 0. \quad (3.19) \]

By letting \( y = 0 \) in (3.19), we get \( \xi^t(x - \overline{x}) \geq 0 \) for all \( x \in S \). From (3.18) it is obvious that

\[ f(x) \geq f(\overline{x}) + \xi^t(x - \overline{x}) \quad \text{for all} \quad x \in \mathbb{R}^n. \]

Therefore, \( \xi \) is a subgradient of \( f \) at \( \overline{x} \) with the property that \( \xi^t(x - \overline{x}) \geq 0 \) for all \( x \in S \), and the proof is complete.

**Corollary 1**

Under the assumptions of Theorem 3.4.3, if \( S \) is open, \( \overline{x} \) is an optimal solution to the problem if and only if there exists a zero subgradient of \( f \) at \( \overline{x} \). In particular, if \( S = \mathbb{R}^n \), \( \overline{x} \) is a global minimum if and only if there exists a zero subgradient of \( f \) at \( \overline{x} \).

**Proof**

By the theorem, \( \overline{x} \) is an optimal solution if and only if \( \xi^t(x - \overline{x}) \geq 0 \) for each \( x \in S \), where \( \xi \) is a subgradient of \( f \) at \( \overline{x} \). Since \( S \) is open, \( x = \overline{x} - \lambda \xi \in S \) for some positive \( \lambda \). Therefore, \( -\lambda \|\xi\|^2 \geq 0 \); that is, \( \xi = 0 \).

**Corollary 2**

In addition to the assumptions of the theorem, suppose that \( f \) is differentiable. Then \( \overline{x} \) is an optimal solution if and only if \( \nabla f(\overline{x})^t(x - \overline{x}) \geq 0 \) for all \( x \in S \). Furthermore, if \( S \) is open, \( \overline{x} \) is an optimal solution if and only if \( \nabla f(\overline{x}) = 0 \).

Note the important implications of Theorem 3.4.3. First, the theorem gives a necessary and sufficient characterization of optimal solutions. This characterization reduces to the well-known condition of vanishing derivatives if \( f \) is differentiable and \( S \) is open. Another important implication is that if we reach a nonoptimal point \( \overline{x} \), where \( \nabla f(\overline{x})^t(x - \overline{x}) < 0 \) for some \( x \in S \), there is an obvious way to proceed to an improving solution. This can be achieved by moving from \( \overline{x} \) in the direction \( d = x - \overline{x} \). The actual size of the step can be determined by solving a line search problem, which is a one-dimensional minimization subproblem of the following form: Minimize \( f[\overline{x} + \lambda d] \) subject to \( \lambda \geq 0 \) and \( \overline{x} + \lambda d \in S \). This procedure is called the method of feasible directions and is discussed in more detail in Chapter 10.

To provide additional insights, let us dwell for awhile on Corollary 2, which addresses the differentiable case for Theorem 3.4.3. Figure 3.8 illustrates
the geometry of the result. Now suppose that for the problem to minimize \( f(x) \) subject to \( x \in S \), we have \( f \) differentiable and convex, but \( S \) is an arbitrary set.

Suppose further that it turns out that the directional derivative \( f'(\bar{x}; x - \bar{x}) = \nabla f(\bar{x})^T (x - \bar{x}) \geq 0 \) for all \( x \in S \). The proof of the theorem actually shows that \( \bar{x} \) is a global minimum regardless of \( S \), since for any solution \( \hat{x} \) that improves over \( \bar{x} \), we have, by the convexity of \( f \), that \( f(\bar{x}) > f(\hat{x}) \geq f(\bar{x}) + \nabla f(\bar{x})^T (\hat{x} - \bar{x}) \), which implies that \( \nabla f(\bar{x})^T (\hat{x} - \bar{x}) < 0 \), whereas \( \nabla f(\bar{x})^T (x - \bar{x}) \geq 0 \) for all \( x \in S \).

Hence, the hyperplane \( \nabla f(\bar{x})^T (x - \bar{x}) = 0 \) separates \( S \) from solutions that improve over \( \bar{x} \). [For the nondifferentiable case, the hyperplane \( \xi^T (x - \bar{x}) = 0 \) plays a similar role.] However, if \( f \) is not convex, the directional derivative \( \nabla f(\bar{x})^T (x - \bar{x}) \) being nonnegative for all \( x \in S \) does not even necessarily imply that \( \bar{x} \) is a local minimum. For example, for the problem to minimize \( f(x) = x^3 \) subject to \(-1 \leq x \leq 1\), we have the condition \( f'(\bar{x})(x - \bar{x}) \geq 0 \) for all \( x \in S \) being satisfied at \( \bar{x} = 0 \), since \( f'(0) = 0 \), but \( \bar{x} = 0 \) is not even a local minimum for this problem.

Conversely, suppose that \( f \) is differentiable but arbitrary otherwise and that \( S \) is a convex set. Then, if \( \bar{x} \) is a global minimum, we must have \( f'(\bar{x}; x - \bar{x}) = \nabla f(\bar{x})^T (x - \bar{x}) \geq 0 \). This follows because, otherwise, if \( \nabla f(\bar{x})^T (x - \bar{x}) < 0 \), we could move along the direction \( d = x - \bar{x} \) and, as above, the objective value would fall for sufficiently small step lengths, whereas \( \bar{x} + \lambda d \) would remain feasible for \( 0 \leq \lambda \leq 1 \) by the convexity of \( S \). Note that this explains a more general concept: if \( f \) is differentiable but \( f \) and \( S \) are otherwise arbitrary, and if \( \bar{x} \) is a local minimum of \( f \) over \( S \), then for any direction \( d \) for which \( \bar{x} + \lambda d \) remains feasible for \( 0 < \lambda \leq \delta \) for some \( \delta > 0 \), we must have a nonnegative
directional derivative of $f$ at $\bar{x}$ in the direction $d$; that is, we must have
$$f'(\bar{x}; d) = \nabla f(\bar{x})^T d \geq 0.$$  

Now let us turn our attention back to convex programming problems. The following result and its corollaries characterize the set of alternative optimal solutions and show, in part, that the gradient of the objective function (assuming twice differentiability) is a constant over the optimal solution set, and that for a quadratic objective function, the optimal solution set is in fact polyhedral. (See Figure 3.8 to identify the set of alternative optimal solutions $S^*$ defined by the theorem in light of Theorem 3.4.3.)

### 3.4.4 Theorem

Consider the problem to minimize $f(x)$ subject to $x \in S$, where $f$ is a convex and twice differentiable function and $S$ is a convex set, and suppose that there exists an optimal solution $\bar{x}$. Then the set of alternative optimal solutions is characterized by the set

$$S^* = \{x \in S : \nabla f(\bar{x})^T (x - \bar{x}) \leq 0 \text{ and } \nabla f(x) = \nabla f(\bar{x})\}.$$  

**Proof**

Denote the set of alternative optimal solutions as $\tilde{S}$, say, and note that $\bar{x} \in \tilde{S} \neq \emptyset$. Consider any $\hat{x} \in S^*$. By the convexity of $f$ and the definition of $S^*$, we have $\hat{x} \in S$ and

$$f(\bar{x}) \geq f(\hat{x}) + \nabla f(\hat{x})^T (\bar{x} - \hat{x}) = f(\hat{x}) + \nabla f(\bar{x})^T (\bar{x} - \hat{x}) \geq f(\hat{x}),$$

so we must have $\hat{x} \in \tilde{S}$ by the optimality of $\bar{x}$. Hence, $S^* \subseteq \tilde{S}$.

Conversely, suppose that $\hat{x} \in \tilde{S}$, so that $\hat{x} \in S$ and $f(\hat{x}) = f(\bar{x})$. This means that $f(\bar{x}) = f(\hat{x}) \geq f(\hat{x}) + \nabla f(\bar{x})^T (\bar{x} - \hat{x})$ or that $\nabla f(\bar{x})^T (\bar{x} - \hat{x}) \leq 0$. But by Corollary 2 to Theorem 3.4.3, we have $\nabla f(\bar{x})^T (\bar{x} - \hat{x}) \geq 0$. Hence, $\nabla f(\bar{x})^T (\bar{x} - \hat{x}) = 0$. By interchanging the roles of $\bar{x}$ and $\hat{x}$, we obtain $\nabla f(\bar{x})^T (\bar{x} - \hat{x}) = 0$ symmetrically. Therefore,

$$[\nabla f(\bar{x}) - \nabla f(\hat{x})]^T (\bar{x} - \hat{x}) = 0. \quad (3.20)$$

Now we have

$$[\nabla f(\bar{x}) - \nabla f(\hat{x})] = \nabla f[\hat{x} + \lambda (\bar{x} - \hat{x})]_{\lambda=0}^{\lambda=1}$$

$$= \int_{\lambda=0}^{\lambda=1} H[\hat{x} + \lambda (\bar{x} - \hat{x})](\bar{x} - \hat{x}) \, d\lambda = G(\bar{x} - \hat{x}), \quad (3.21)$$

where

$$H[\hat{x} + \lambda (\bar{x} - \hat{x})]$$. 


where \( G = \int_0^1 H[\hat{x} + \lambda(\bar{x} - \hat{x})] \, d\lambda \) and where the integral of the matrix is performed componentwise. But note that \( G \) is positive semidefinite because \( d'Gd = \int_0^1 d'H[\hat{x} + \lambda(\bar{x} - \hat{x})]d \, d\lambda \geq 0 \) for all \( d \in \mathbb{R}^n \), since \( d'H[\hat{x} + \lambda(\bar{x} - \hat{x})]d \) is a non-negative function of \( \lambda \) by the convexity of \( f \). Hence, by (3.20) and (3.21), we get

\[
0 = (\bar{x} - \hat{x})'[\nabla f(\bar{x}) - \nabla f(\hat{x})] = (\bar{x} - \hat{x})'G(\bar{x} - \hat{x}).
\]

But the positive semidefiniteness of \( G \) implies that \( C(X - \bar{x}) = 0 \) by a standard result (see Exercise 3.41). Therefore, by (3.21), we have \( \nabla f(\bar{x}) = \nabla f(\hat{x}) \). We have hence shown that \( \hat{x} \in S \), \( \nabla f(\bar{x})' (\hat{x} - \bar{x}) \leq 0 \), and \( \nabla f(\hat{x}) = \nabla f(\bar{x}) \). This means that \( \hat{x} \in S^* \), and thus \( \bar{S} \subseteq S^* \).

This, together with \( S^* \subseteq \bar{S} \), completes the proof.

**Corollary 1**

The set \( S^* \) of alternative optimal solutions can equivalently be defined as

\[
S^* = \{ x \in S : \nabla f(\bar{x})' (x - \bar{x}) = 0 \text{ and } \nabla f(x) = \nabla f(\bar{x}) \}.
\]

**Proof**

The proof follows from the definition of \( S^* \) in Theorem 3.4.4 and the fact that \( \nabla f(\bar{x})' (x - \bar{x}) \geq 0 \) for all \( x \in S \) by Corollary 2 to Theorem 3.4.3.

**Corollary 2**

Suppose that \( f \) is a quadratic function given by \( f(x) = c'x + (1/2)x'Hx \) and that \( S \) is polyhedral. Then \( S^* \) is a polyhedral set given by

\[
S^* = \{ x \in S : c'(X - \bar{x}) \leq 0, \ H(x - \bar{x}) = 0 \} = \{ x \in S : c'(X - \bar{x}) = 0, \ H(x - \bar{x}) = 0 \}.
\]

**Proof**

The proof follows by direct substitution in Theorem 3.4.4 and Corollary 1, noting that \( \nabla f(x) = c + Hx \).
### 3.4.5 Example

Minimize \( \left(x_1 - \frac{3}{2}\right)^2 + (x_2 - 5)^2 \)

subject to

\[
\begin{align*}
-x_1 + x_2 & \leq 2 \\
2x_1 + 3x_2 & \leq 11 \\
-x_1 & \leq 0 \\
-x_2 & \leq 0.
\end{align*}
\]

Clearly, \( f(x_1, x_2) = \left(x_1 - \frac{3}{2}\right)^2 + (x_2 - 5)^2 \) is a convex function, which gives the square of the distance from the point \((3/2, 5)\). The convex polyhedral set \( S \) is represented by the above four inequalities. The problem is depicted in Figure 3.9. From the figure, clearly the optimal point is \((1, 3)\). The gradient vector of \( f \) at the point \((1, 3)\) is \( \nabla f(1,3) = (-1, -4)' \). We see geometrically that the vector \((-1, -4)\) makes an angle of \(< 90^\circ\) with each vector of the form \((x_1 - 1, x_2 - 3)\), where \((x_1, x_2) \in S\). Thus, the optimality condition of Theorem 3.4.3 is verified and, by Theorem 3.4.4, \((1, 3)\) is the unique optimum.

To illustrate further, suppose that it is claimed that \( \hat{x} = (0, 0)' \) is an optimal point. By Theorem 3.4.4, this cannot be true since we have \( \nabla f(\overline{x})'(\hat{x} - \overline{x}) = 13 > 0 \) when \( \overline{x} = (1, 3)' \). Similarly, by Theorem 3.4.3, we can easily verify that \( \hat{x} \) is not optimal. Note that \( \nabla f(0,0) = (-3, -10)' \) and actually, for each nonzero \( x \in S \), we have \(-3x_1 - 10x_2 < 0\). Hence, the origin could not be an optimal point. Moreover, we can improve \( f \) by moving from \( 0 \) in the direction \( x - 0 \) for any \( x \in S \). In this case, the best local direction is \(-\nabla f(0,0)\), that is, the direction \((3, 10)\). In Chapter 10 we discuss methods for finding a particular direction among many alternatives.

---

**Figure 3.9 Setup for Example 3.4.5.**
Maximizing a Convex Function

We now develop a necessary condition for a maximum of a convex function over a convex set. Unfortunately, this condition is not sufficient. Therefore, it is possible, and actually not unlikely, that several local maxima satisfying the condition of Theorem 3.4.6 exist. Unlike the minimization case, there exists no local information at such solutions that could lead us to better points. Hence, maximizing a convex function is usually a much harder task than minimizing a convex function. Again, minimizing a concave function is similar to maximizing a convex function, and hence the development for the concave case is left to the reader.

3.4.6 Theorem

Let \( f: \mathbb{R}^n \to \mathbb{R} \) be a convex function, and let \( S \) be a nonempty convex set in \( \mathbb{R}^n \). Consider the problem to maximize \( f(x) \) subject to \( x \in S \). If \( \bar{x} \in S \) is a local optimal solution, \( \xi^\ell(x - \bar{x}) \leq 0 \) for each \( x \in S \), where \( \xi \) is any subgradient of \( f \) at \( \bar{x} \).

Proof

Suppose that \( \bar{x} \in S \) is a local optimal solution. Then an \( \varepsilon \)-neighborhood \( N_\varepsilon(\bar{x}) \) exists such that \( f(x) \leq f(\bar{x}) \) for each \( x \in S \cap N_\varepsilon(\bar{x}) \). Let \( x \in S \), and note that \( \bar{x} + \lambda(x - \bar{x}) \in S \cap N_\varepsilon(\bar{x}) \) for \( \lambda > 0 \) and sufficiently small. Hence,

\[
f[\bar{x} + \lambda(x - \bar{x})] \leq f(\bar{x}).
\]

(3.22)

Let \( \xi \) be a subgradient of \( f \) at \( \bar{x} \). By the convexity of \( f \), we have

\[
f[\bar{x} + \lambda(x - \bar{x})] - f(\bar{x}) \geq \lambda \xi^\ell(x - \bar{x}).
\]

The above inequality, together with (3.20), implies that \( \lambda \xi^\ell(x - \bar{x}) \leq 0 \), and dividing by \( \lambda > 0 \), the result follows.

Corollary

In addition to the assumptions of the theorem, suppose that \( f \) is differentiable. If \( \bar{x} \in S \) is a local optimal solution, \( \nabla f(\bar{x})^\ell(x - \bar{x}) \leq 0 \) for all \( x \in S \).

Note that the above result is, in general, necessary but not sufficient for optimality. To illustrate, let \( f(x) = x^2 \) and \( S = \{x : -1 \leq x \leq 2\} \). The maximum of \( f \) over \( S \) is equal to 4 and is achieved at \( x = 2 \). However, at \( \bar{x} = 0 \), we have \( \nabla f(\bar{x}) = 0 \) and hence \( \nabla f(\bar{x})^\ell(x - \bar{x}) = 0 \) for each \( x \in S \). Clearly, the point \( \bar{x} = 0 \) is not even a local maximum. Referring to Example 3.4.5, discussed earlier, we
have two local maxima, (0, 0) and (1½, 0). Both points satisfy the necessary condition of Theorem 3.4.6. If we are currently at the local optimal point (0, 0), unfortunately no local information exists that will lead us toward the global maximum point (1½, 0). Also, if we are at the global maximum point (1½, 0), there is no convenient local criterion that tells us that we are at the optimal point.

Theorem 3.4.7 shows that a convex function achieves a maximum over a compact polyhedral set at an extreme point. This result has been utilized by several computational schemes for solving such problems. We ask the reader to think for a moment about the case when the objective function is linear and, hence, both convex and concave. Theorem 3.4.7 could be extended to the case where the convex feasible region is not polyhedral.

3.4.7 Theorem

Let \( f: \mathbb{R}^n \to \mathbb{R} \) be a convex function, and let \( S \) be a nonempty compact polyhedral set in \( \mathbb{R}^n \). Consider the problem to maximize \( f(x) \) subject to \( x \in S \). An optimal solution \( \bar{x} \) to the problem then exists, where \( \bar{x} \) is an extreme point of \( S \).

Proof

By Theorem 3.1.3, note that \( f \) is continuous. Since \( S \) is compact, \( f \) assumes a maximum at \( x' \in S \). If \( x' \) is an extreme point of \( S \), the result is at hand. Otherwise, by Theorem 2.6.7, \( x' = \sum_{j=1}^{k} \lambda_j x_j \), where \( \sum_{j=1}^{k} \lambda_j = 1 \), \( \lambda_j > 0 \), and \( x_j \) is an extreme point of \( S \) for \( j = 1, \ldots, k \). By the convexity of \( f \), we have

\[
  f(x') = f \left( \sum_{j=1}^{k} \lambda_j x_j \right) \leq \sum_{j=1}^{k} \lambda_j f(x_j).
\]

But since \( f(x') \geq f(x_j) \) for \( j = 1, \ldots, k \), the above inequality implies that \( f(x') = f(x_j) \) for \( j = 1, \ldots, k \). Thus, the extreme points \( x_1, \ldots, x_k \) are optimal solutions to the problem, and the proof is complete.

3.5 Generalizations of a Convex Functions

In this section we present various types of functions that are similar to convex and concave functions but that share only some of their desirable properties. As we shall learn, many of the results presented later in the book do not require the restrictive assumption of convexity, but rather, the less restrictive assumptions of quasiconvexity, pseudoconvexity, and convexity at a point.

Quasiconvex Functions

Definition 3.5.1 introduces quasiconvex functions. From the definition it is apparent that every convex function is also quasiconvex.
### 3.5.1 Definition

Let $f: S \to R$, where $S$ is a nonempty convex set in $R^n$. The function $f$ is said to be \textit{quasiconvex} if for each $x_1$ and $x_2 \in S$, the following inequality is true:

$$f[\lambda x_1 + (1-\lambda)x_2] \leq \max\{f(x_1), f(x_2)\} \text{ for each } \lambda \in (0, 1).$$

The function $f$ is said to be \textit{quasiconcave} if $-f$ is quasiconvex.

From Definition 3.5.1, a function $f$ is quasiconvex if whenever $f(x_2) \geq f(x_1)$, $f(x_2)$ is greater than or equal to $f$ at all convex combinations of $x_1$ and $x_2$. Hence, if $f$ increases from its value at a point along any direction, it must remain nondecreasing in that direction. Therefore, its univariate cross section is either monotone or unimodal (see Exercise 3.57). A function $f$ is quasiconcave if whenever $f(x_2) \geq f(x_1)$, $f$ at all convex combinations of $x_1$ and $x_2$ is greater than or equal to $f(x_1)$. Figure 3.10 shows some examples of quasiconvex and quasiconcave functions. We shall concentrate on quasiconvex functions; the reader is advised to draw all the parallel results for quasiconcave functions. A function that is both quasiconvex and quasiconcave is called \textit{quasimonotone} (see Figure 3.10d).

We have learned in Section 3.2 that a convex function can be characterized by the convexity of its epigraph. We now learn that a quasiconvex function can be characterized by the convexity of its level sets. This result is given in Theorem 3.5.2.

### 3.5.2 Theorem

Let $f: S \to R$ where $S$ is a nonempty convex set in $R^n$. The function $f$ is quasiconvex if and only if $S_\alpha = \{x \in S : f(x) \leq \alpha\}$ is convex for each real number $\alpha$. 

---

**Figure 3.10 Quasiconvex and quasiconcave functions:** (a) quasiconvex, (b) quasiconcave, (c) neither quasiconvex nor quasiconcave, (d) quasimonotone.
Suppose that $f$ is quasiconvex, and let $x_1, x_2 \in S$. Therefore, $x_1, x_2 \in S$ and $\max\{f(x_1), f(x_2)\} \leq \alpha$. Let $\lambda \in (0,1)$, and let $x = \lambda x_1 + (1-\lambda)x_2$. By the convexity of $S$, $x \in S$. Furthermore, by the quasiconvexity of $f$, $f(x) \leq \max\{f(x_1), f(x_2)\} \leq \alpha$. Hence, $x \in S_\alpha$ and thus $S_\alpha$ is convex. Conversely, suppose that $S_\alpha$ is convex for each real number $\alpha$. Let $x_1, x_2 \in S$. Furthermore, let $\lambda \in (0,1)$ and $x = \lambda x_1 + (1-\lambda)x_2$. Note that $x_1, x_2 \in S_\alpha$ for $\alpha = \max\{f(x_1), f(x_2)\}$. By assumption, $S_\alpha$ is convex, so that $x \in S_\alpha$. Therefore, $f(x) \leq \alpha = \max\{f(x_1), f(x_2)\}$. Hence, $f$ is quasiconvex, and the proof is complete.

The level set $S_\alpha$ defined in Theorem 3.5.2 is sometimes referred to as a lower-level set, to differentiate it from the upper-level set $\{x \in S : f(x) \geq \alpha\}$, which is convex for all $\alpha \in R$ if and only if $f$ is quasiconcave. Also, it can be shown (see Exercise 3.59) that $f$ is quasimonotone if and only if the level surface $\{x \in S : f(x) = \alpha\}$ is convex for all $\alpha \in R$.

We now give a result analogous to Theorem 3.4.7. Theorem 3.5.3 shows that the maximum of a continuous quasiconvex function over a compact polyhedral set occurs at an extreme point.

3.5.3 Theorem

Let $S$ be a nonempty compact polyhedral set in $R^n$, and let $f: R^n \to R$ be quasiconvex and continuous on $S$. Consider the problem to maximize $f(x)$ subject to $x \in S$. Then an optimal solution $\bar{x}$ to the problem exists, where $\bar{x}$ is an extreme point of $S$.

Proof

Note that $f$ is continuous on $S$ and hence attains a maximum, say, at $x' \in S$. If there is an extreme point whose objective is equal to $f(x')$, the result is at hand. Otherwise, let $x_1, \ldots, x_k$ be the extreme points of $S$, and assume that $f(x') > f(x_j)$ for $j = 1, \ldots, k$. By Theorem 2.6.7, $x'$ can be represented as

$$x' = \sum_{j=1}^{k} \lambda_j x_j$$

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

$$\lambda_j \geq 0, \quad j = 1, \ldots, k.$$

Since $f(x') > f(x_j)$ for each $j$, then
Now, consider the set $S_\alpha = \{x : f(x) \leq \alpha\}$. Note that $x_j \in S_\alpha$ for $j = 1, \ldots, k$, and by the quasiconvexity of $f$, $S_\alpha$ is convex. Hence, $x' = \sum_{j=1}^{k} \lambda_j x_j$ belongs to $S_\alpha$. This implies that $f(x') \leq \alpha$, which contradicts (3.23). This contradiction shows that $f(x') = f(x_j)$ for some extreme point $x_j$, and the proof is complete.

### Differentiable Quasiconvex Functions

The following theorem gives a necessary and sufficient characterization of a differentiable quasiconvex function. (See Appendix B for a second-order characterization in terms of bordered Hessian determinants.)

**3.5.4 Theorem**

Let $S$ be a nonempty open convex set in $\mathbb{R}^n$, and let $f: S \to \mathbb{R}$ be differentiable on $S$. Then $f$ is quasiconvex if and only if either one of the following equivalent statements holds true:

1. If $x_1, x_2 \in S$ and $f(x_1) \leq f(x_2)$, $\nabla f(x_2) \cdot (x_1 - x_2) \leq 0$.
2. If $x_1, x_2 \in S$ and $\nabla f(x_2) \cdot (x_1 - x_2) > 0$, $f(x_1) > f(x_2)$.

**Proof**

Obviously, statements 1 and 2 are equivalent. We shall prove Part 1. Let $f$ be quasiconvex, and let $x_1, x_2 \in S$ be such that $f(x_1) < f(x_2)$. By the differentiability of $f$ at $x_2$, for $\lambda \in (0, 1)$, we have

$$f(\lambda x_1 + (1-\lambda) x_2) - f(x_2) = \lambda \nabla f(x_2)^\top (x_1 - x_2) + \lambda \|x_1 - x_2\| \alpha[x_2; \lambda(x_1 - x_2)],$$

where $\alpha[x_2; \lambda(x_1 - x_2)] \to 0$ as $\lambda \to 0$. By the quasiconvexity of $f$, we have $f(\lambda x_1 + (1-\lambda) x_2) \leq f(x_2)$, and hence the above equation implies that

$$\lambda \nabla f(x_2)^\top (x_1 - x_2) + \lambda \|x_1 - x_2\| \alpha[x_2; \lambda(x_1 - x_2)] \leq 0.$$

Dividing by $\lambda$ and letting $\lambda \to 0$, we get $\nabla f(x_2)^\top (x_1 - x_2) \leq 0$.

Conversely, suppose that $x_1, x_2 \in S$ and that $f(x_1) \leq f(x_2)$. We need to show that given Part 1, we have $f(\lambda x_1 + (1-\lambda) x_2) \leq f(x_2)$ for each $\lambda \in (0, 1)$. We do this by showing that the set

$$L = \{x : x = \lambda x_1 + (1-\lambda) x_2, \lambda \in (0,1), f(x) > f(x_2)\}$$
is empty. By contradiction, suppose that there exists an $x' \in L$. Therefore, $x' = \lambda x_1 + (1 - \lambda) x_2$ for some $\lambda \in (0, 1)$ and $f(x') > f(x_2)$. Since $f$ is differentiable, it is continuous, and there must exist a $\delta \in (0, 1)$ such that

$$f(\mu x' + (1 - \mu) x_2) > f(x_2) \quad \text{for each } \mu \in [\delta, 1]$$

and $f(x') > f(\delta x' + (1 - \delta) x_2)$. By this inequality and the mean value theorem, we must have

$$0 < f(x') - f(\delta x' + (1 - \delta) x_2) = (1 - \delta) \nabla f(\hat{x})^T (x' - x_2),$$

where $\hat{x} = \mu x' + (1 - \mu) x_2$ for some $\mu \in (\delta, 1)$. From (3.24) it is clear that $f(\hat{x}) > f(x_2)$. Dividing (3.25) by $1 - \delta > 0$, it follows that $\nabla f(\hat{x})^T (x' - x_2) > 0$, which in turn implies that

$$\nabla f(\hat{x})^T (x_1 - x_2) > 0.$$  

But on the other hand, $f(\hat{x}) \geq f(x_2) \geq f(x_1)$, and $\hat{x}$ is a convex combination of $x_1$ and $x_2$, say $\hat{x} = \hat{\lambda} x_1 + (1 - \hat{\lambda}) x_2$, where $\hat{\lambda} \in (0, 1)$. By the assumption of the theorem, $\nabla f(\hat{x})^T (x_1 - \hat{x}) \leq 0$, and thus we must have

$$0 \geq \nabla f(\hat{x})^T (x_1 - \hat{x}) = (1 - \hat{\lambda}) \nabla f(\hat{x})^T (x_1 - x_2).$$

The above inequality is not compatible with (3.26). Therefore, $L$ is empty, and the proof is complete.

To illustrate Theorem 3.5.4, let $f(x) = x^3$. To check its quasiconvexity, suppose that $f(x_1) \leq f(x_2)$, that is, $x_1^3 \leq x_2^3$. This is true only if $x_1 \leq x_2$. Now consider $\nabla f(x_2)(x_1 - x_2) = 3(x_1 - x_2)x_2^2$. Since $x_1 \leq x_2$, $3(x_1 - x_2)x_2^2 \leq 0$. Therefore, $f(x_1) \leq f(x_2)$ implies that $\nabla f(x_2)(x_1 - x_2) \leq 0$, and by the theorem we have that $f$ is quasiconvex. As another illustration, let $f(x_1, x_2) = x_1^3 + x_2^3$. Let $x_1 = (2, -2)^T$ and $x_2 = (1, 0)^T$. Note that $f(x_1) = 0$ and $f(x_2) = 1$, so that $f(x_1) < f(x_2)$. But on the other hand, $\nabla f(x_2)^T (x_1 - x_2) = (3, 0)(1, -2)^T = 3$. By the necessary part of the theorem, $f$ is not quasiconvex. This also shows that the sum of two quasiconvex functions is not necessarily quasiconvex.

**Strictly Quasiconvex Functions**

Strictly quasiconvex and strictly quasiconcave functions are especially important in nonlinear programming because they ensure that a local minimum and a local
maximum over a convex set are, respectively, a global minimum and a global maximum.

3.5.5 Definition

Let \( f: S \to \mathbb{R} \), where \( S \) is a nonempty convex set in \( \mathbb{R}^n \). The function \( f \) is said to be strictly quasiconvex if for each \( x_1, x_2 \in S \) with \( f(x_1) \neq f(x_2) \), we have

\[
f(\lambda x_1 + (1 - \lambda)x_2) < \max\{f(x_1), f(x_2)\} \quad \text{for each } \lambda \in (0, 1).
\]

The function \( f \) is called strictly quasiconcave if \(-f\) is strictly quasiconvex. Strictly quasiconvex functions are also sometimes referred to as semi-strictly quasiconvex, functionally convex, or explicitly quasiconvex.

Note from Definition 3.5.5 that every convex function is strictly quasiconvex. Figure 3.11 gives examples of strictly quasiconvex and strictly quasiconcave functions. Also, the definition precludes any "flat spots" from occurring anywhere except at extremizing points. This is formalized by the following theorem, which shows that a local minimum of a strictly quasiconvex function over a convex set is also a global minimum. This property is not enjoyed by quasiconvex functions, as seen in Figure 3.10a.

3.5.6 Theorem

Let \( f: \mathbb{R}^n \to \mathbb{R} \) be strictly quasiconvex. Consider the problem to minimize \( f(x) \) subject to \( x \in S \), where \( S \) is a nonempty convex set in \( \mathbb{R}^n \). If \( \bar{x} \) is a local optimal solution, \( \bar{x} \) is also a global optimal solution.

**Proof**

Assume, on the contrary, that there exists an \( \hat{x} \in S \) with \( f(\hat{x}) < f(\bar{x}) \). By the convexity of \( S \), \( \lambda \hat{x} + (1 - \lambda)\bar{x} \in S \) for each \( \lambda \in (0, 1) \). Since \( \bar{x} \) is a local minimum by assumption, \( f(\bar{x}) \leq f(\lambda \hat{x} + (1 - \lambda)\bar{x}) \) for all \( \lambda \in (0, \delta) \) and for some \( \delta \in \).
But because $f$ is strictly quasiconvex and $f(\bar{x}) < f(x)$, we have that $f(\lambda x + (1 - \lambda)\bar{x}) < f(\bar{x})$ for each $\lambda \in (0, 1)$. This contradicts the local optimality of $\bar{x}$, and the proof is complete.

As seen from Definition 3.1.1, every strictly convex function is indeed a convex function. But every strictly quasiconvex function is not quasiconvex. To illustrate, consider the following function given by Karamardian [1967]:

$$f(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \neq 0. \end{cases}$$

By Definition 3.5.5, $f$ is strictly quasiconvex. However, $f$ is not quasiconvex, since for $x_1 = 1$ and $x_2 = -1$, $f(x_1) = f(x_2) = 0$, but $f[(1/2)x_1 + (1/2)x_2] = f(0) = 1 > f(x_2)$. If $f$ is lower semicontinuous, however, then as shown below, strict quasiconvexity implies quasiconvexity, as one would usually expect from the word strict. (For a definition of lower semicontinuity, refer to Appendix A.)

3.5.7 Lemma

Let $S$ be a nonempty convex set in $\mathbb{R}^n$ and let $f: S \to \mathbb{R}$ be strictly quasiconvex and lower semicontinuous. Then $f$ is quasiconvex.

Proof

Let $x_1$ and $x_2 \in S$. If $f(x_1) \neq f(x_2)$, then by the strict quasiconvexity of $f$, we must have $f(\lambda x_1 + (1 - \lambda)x_2) < \max\{f(x_1), f(x_2)\}$ for each $\lambda \in (0, 1)$. Now, suppose that $f(x_1) = f(x_2)$. To show that $f$ is quasiconvex, we need to show that $f(\lambda x_1 + (1 - \lambda)x_2) < f(x_1)$ for each $\lambda \in (0, 1)$. By contradiction, suppose that $f(\mu x_1 + (1 - \mu)x_2) > f(x_1)$ for some $\mu \in (0, 1)$. Denote $\mu x_1 + (1 - \mu)x_2$ by $x$. Since $f$ is lower semicontinuous, there exists a $\lambda \in (0, 1)$ such that

$$f(x) > f(\lambda x_1 + (1 - \lambda)x) > f(x_1) = f(x_2). \quad (3.27)$$

Note that $x$ can be represented as a convex combination of $\lambda x_1 + (1 - \lambda)x$ and $x_2$. Hence, by the strict quasiconvexity of $f$ and since $f(\lambda x_1 + (1 - \lambda)x) > f(x_2)$, we have $f(x) < f(\lambda x_1 + (1 - \lambda)x) < f(x_2)$, contradicting (3.27). This completes the proof.

Strongly Quasiconvex Functions

From Theorem 3.5.6 it followed that a local minimum of a strictly quasiconvex function over a convex set is also a global optimal solution. However, strict quasiconvexity does not assert uniqueness of the global optimal solution. We shall define here another version of quasiconvexity, called strong quasiconvexity, which assures uniqueness of the global minimum when it exists.
3.5.8 Definition

Let $S$ be a nonempty convex set in $\mathbb{R}^n$, and let $f: S \rightarrow \mathbb{R}$. The function $f$ is said to be **strongly quasiconvex** if for each $x_1, x_2 \in S$, with $x_1 \neq x_2$, we have

$$f(\lambda x_1 + (1-\lambda)x_2) < \max\{f(x_1), f(x_2)\}$$

for each $\lambda \in (0, 1)$. The function $f$ is said to be **strongly quasiconcave** if $-f$ is strongly quasiconvex. (We caution the reader that such a function is sometimes referred to in the literature as being *strictly quasiconvex*, whereas a function satisfying Definition 3.5.5 is called *semi-strictly quasiconvex*. This is done because of Karamardian's example given above and Property 3 below.)

From Definition 3.5.8 and from Definitions 3.1.1, 3.5.1, and 3.5.5, the following statements hold true:

1. Every strictly convex function is strongly quasiconvex.
2. Every strongly quasiconvex function is strictly quasiconvex.
3. Every strongly quasiconvex function is quasiconvex even in the absence of any semicontinuity assumption.

Figure 3.11a illustrates a case where the function is both strongly quasiconvex and strictly quasiconvex, whereas the function represented in Figure 3.11b is strictly quasiconvex but not strongly quasiconvex. The key to strong quasiconvexity is that it enforces strict unimodality (see Exercise 3.58). This leads to the following property.

3.5.9 Theorem

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be strongly quasiconvex. Consider the problem to minimize $f(x)$ subject to $x \in S$, where $S$ is a nonempty convex set in $\mathbb{R}^n$. If $\bar{x}$ is a local optimal solution, $\bar{x}$ is the unique global optimal solution.

**Proof**

Since $\bar{x}$ is a local optimal solution, there exists an $\varepsilon$-neighborhood $N_\varepsilon(\bar{x})$ around $\bar{x}$ such that $f(\bar{x}) \leq f(x)$ for all $x \in S \cap N_\varepsilon(\bar{x})$. Suppose, by contradiction to the conclusion of the theorem, that there exists a point $\hat{x} \in S$ such that $\hat{x} \neq \bar{x}$ and $f(\hat{x}) \leq f(\bar{x})$. By strong quasiconvexity it follows that

$$f(\lambda \hat{x} + (1-\lambda)\bar{x}) < \max\{f(\hat{x}), f(\bar{x})\} = f(\bar{x})$$

for all $\lambda \in (0, 1)$. But for $\lambda$ small enough, $\lambda \hat{x} + (1-\lambda)\bar{x} \in S \cap N_\varepsilon(\bar{x})$, so that the above inequality violates the local optimality of $\bar{x}$. This completes the proof.

Pseudoconvex Functions

The astute reader might already have observed that differentiable strongly (or strictly) quasiconvex functions do not share the particular property of convex
functions, which says that if $\nabla f(\bar{x}) = 0$ at some point $\bar{x}$, $\bar{x}$ is a global minimum of $f$. Figure 3.12c illustrates this fact. This motivates the definition of pseudoconvex functions that share this important property with convex functions, and leads to a generalization of various derivative-based optimality conditions.

### 3.5.10 Definition

Let $S$ be a nonempty open set in $\mathbb{R}^n$, and let $f: S \to \mathbb{R}$ be differentiable on $S$. The function $f$ is said to be **pseudoconvex** if for each $x_1, x_2 \in S$ with $\nabla f(x_1)^T (x_2 - x_1) \geq 0$, we have $f(x_2) \geq f(x_1)$; or equivalently, if $f(x_2) < f(x_1)$, $\nabla f(x_1)^T (x_2 - x_1) < 0$. The function $f$ is said to be **pseudoconcave** if $-f$ is pseudoconvex.

The function $f$ is said to be **strictly pseudoconvex** if for each distinct $x_1, x_2 \in S$ satisfying $\nabla f(x_1)^T (x_2 - x_1) \geq 0$, we have $f(x_2) \geq f(x_1)$; or equivalently, if for each distinct $x_1, x_2 \in S$, $f(x_2) \leq f(x_1)$ implies that $\nabla f(x_1)^T (x_2 - x_1) < 0$. The function $f$ is said to be **strictly pseudoconcave** if $-f$ is strictly pseudoconvex.

Figure 3.12a illustrates a pseudoconvex function. From the definition of pseudoconvexity it is clear that if $\nabla f(\bar{x}) = 0$ at any point $\bar{x}$, $f(x) \geq f(\bar{x})$ for all $x$; so $\bar{x}$ is a global minimum for $f$. Hence, the function in Figure 3.12c is neither pseudoconvex nor pseudoconcave. In fact, the definition asserts that if the directional derivative of $f$ at any point $x_1$ in the direction $(x_2 - x_1)$ is nonnegative, the function values are nondecreasing in that direction (see Exercise 3.69). Furthermore, observe that the pseudoconvex functions shown in Figure 3.12 are also strictly quasiconvex, which is true in general, as shown by Theorem 3.5.11. The reader may note that the function in Figure 3.8c is not pseudoconvex, yet it is strictly quasiconvex.

---

**Figure 3.12** Pseudoconvex and pseudoconcave functions: (a) pseudoconvex, (b) both pseudoconvex and pseudoconcave, (c) neither pseudoconvex nor pseudoconcave.
3.5.11 Theorem

Let $S$ be a nonempty open convex set in $\mathbb{R}^n$, and let $f: S \rightarrow \mathbb{R}$ be a differentiable pseudoconvex function on $S$. Then $f$ is both strictly quasiconvex and quasiconvex.

**Proof**

We first show that $f$ is strictly quasiconvex. By contradiction, suppose that there exist $x_1, x_2 \in S$ such that $f(x_1) \neq f(x_2)$ and $f(x') \geq \max\{f(x_1), f(x_2)\}$, where $x' = \lambda x_1 + (1 - \lambda)x_2$ for some $\lambda \in (0, 1)$. Without loss of generality, assume that $f(x_1) < f(x_2)$, so that

$$f(x') \geq f(x_2) > f(x_1).$$  \hspace{1cm} (3.28)

Note, by the pseudoconvexity of $f$, that $\nabla f(x')'(x_1 - x') < 0$. Now since $\nabla f(x')'(x_1 - x') < 0$ and $x_1 - x' = -(1 - \lambda)(x_2 - x')/\lambda$, $\nabla f(x')'(x_2 - x') > 0$; and hence, by the pseudoconvexity of $f$, we must have $f(x_2) \geq f(x')$. Therefore, by (3.28), we get $f(x_2) = f(x')$. Also, since $\nabla f(x')'(x_2 - x') > 0$, there exists a point $\hat{x} = \mu x' + (1 - \mu)x_2$ with $\mu \in (0, 1)$ such that

$$f(\hat{x}) > f(x') = f(x_2).$$

Again, by the pseudoconvexity of $f$, we have $\nabla f(\hat{x})'(x_2 - \hat{x}) < 0$. Similarly, $\nabla f(\hat{x})'(x' - \hat{x}) < 0$. Summarizing, we must have

$$\nabla f(\hat{x})'(x_2 - \hat{x}) < 0$$

$$\nabla f(\hat{x})'(x' - \hat{x}) < 0.$$

Note that $x_2 - \hat{x} = \mu(\hat{x} - x')/(1 - \mu)$, and hence the above two inequalities are not compatible. This contradiction shows that $f$ is strictly quasiconvex. By Lemma 3.5.7, then $f$ is also quasiconvex, and the proof is complete.

In Theorem 3.5.12 we see that every strictly pseudoconvex function is strongly quasiconvex.

3.5.12 Theorem

Let $S$ be a nonempty open convex set in $\mathbb{R}^n$, and let $f: S \rightarrow \mathbb{R}$ be a differentiable strictly pseudoconvex function. Then $f$ is strongly quasiconvex.

**Proof**

By contradiction, suppose that there exist distinct $x_1, x_2 \in S$ and $\lambda \in (0, 1)$ such that $f(x) \geq \max\{f(x_1), f(x_2)\}$, where $x = \lambda x_1 + (1 - \lambda)x_2$. Since $f(x_1)$
\[ f(x), \text{ we have, by the strict pseudoconvexity of } f, \text{ that } \nabla f(x)'(x_1 - x) < 0 \text{ and hence} \]

\[ \nabla f(x)'(x_1 - x_2) < 0. \quad (3.29) \]

Similarly, since \( f(x_2) \leq f(x) \), we have

\[ \nabla f(x)'(x_2 - x_1) < 0. \quad (3.30) \]

The two inequalities (3.29) and (3.30) are not compatible, and hence \( f \) is strongly quasiconvex. This completes the proof.

We remark here in connection with Theorems 3.5.11 and 3.5.12, for the special case in which \( f \) is quadratic, that \( f \) is pseudoconvex if and only if \( f \) is strictly quasiconvex, which holds true if and only if \( f \) is quasiconvex. Moreover, we also have that \( f \) is strictly pseudoconvex if and only if \( f \) is strongly quasiconvex. Hence, all these properties become equivalent to each other for quadratic functions (see Exercise 3.55). Also, Appendix B provides a bordered Hessian determinant characterization for checking the pseudoconvexity and the strict pseudoconvexity of quadratic functions.

Thus far we have discussed various types of convexity and concavity. Figure 3.13 summarizes the implications among these types of convexity. These implications either follow from the definitions or from the various results proved in this section. A similar figure can be constructed for the concave case.

Figure 3.13 Relationship among various types of convexity.
Convexity at a Point

Another useful concept in optimization is the notion of convexity or concavity at a point. In some cases the requirement of a convex or concave function may be too strong and really not essential. Instead, convexity or concavity at a point may be all that is needed.

3.5.13 Definition

Let $S$ be a nonempty convex set in $\mathbb{R}^n$, and let $f: S \to \mathbb{R}$. The following are relaxations of various forms of convexity presented in this chapter:

**Convexity at $\bar{x}$.** The function $f$ is said to be convex at $\bar{x} \in S$ if

\[ f[\lambda \bar{x} + (1-\lambda)x] \leq \lambda f(\bar{x}) + (1-\lambda)f(x) \]

for each $\lambda \in (0, 1)$ and each $x \in S$.

**Strict convexity at $\bar{x}$.** The function $f$ is said to be strictly convex at $\bar{x} \in S$ if

\[ f[\lambda \bar{x} + (1-\lambda)x] < \lambda f(\bar{x}) + (1-\lambda)f(x) \]

for each $\lambda \in (0, 1)$ and for each $x \in S$, $x \neq \bar{x}$.

**Quasiconvexity at $\bar{x}$.** The function $f$ is said to be quasiconvex at $\bar{x} \in S$ if

\[ f[\lambda \bar{x} + (1-\lambda)x] \leq \max\{f(x), f(\bar{x})\} \]

for each $\lambda \in (0, 1)$ and each $x \in S$.

**Strict quasiconvexity at $\bar{x}$.** The function $f$ is said to be strictly quasiconvex at $\bar{x} \in S$ if

\[ f[\lambda \bar{x} + (1-\lambda)x] < \max\{f(x), f(\bar{x})\} \]

for each $\lambda \in (0, 1)$ and each $x \in S$ such that $f(x) \neq f(\bar{x})$.

**Strong quasiconvexity at $\bar{x}$.** The function $f$ is said to be strongly quasiconvex at $\bar{x} \in S$ if

\[ f[\lambda \bar{x} + (1-\lambda)x] < \max\{f(x), f(\bar{x})\} \]

for each $\lambda \in (0, 1)$ and each $x \in S$, $x \neq \bar{x}$.

**Pseudoconvexity at $\bar{x}$.** The function $f$ is said to be pseudoconvex at $\bar{x} \in S$ if $\nabla f(\bar{x})^T(x - \bar{x}) \geq 0$ for $x \in S$ implies that $f(x) \geq f(\bar{x})$.

**Strict pseudoconvexity at $\bar{x}$.** The function $f$ is said to be strictly pseudoconvex at $\bar{x} \in S$ if $\nabla f(\bar{x})^T(x - \bar{x}) \geq 0$ for $x \in S$, $x \neq \bar{x}$, implies that $f(x) > f(\bar{x})$. 
Various types of concavity at a point can be stated in a similar fashion. Figure 3.14 shows some types of convexity at a point. As the figure suggests, these types of convexity at a point represent a significant relaxation of the concept of convexity.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{convexity_types.png}
\caption{Various types of convexity at a point. (a) Convexity and strict convexity: \( f \) is convex but not strictly convex at \( x_1 \); \( f \) is both convex and strictly convex at \( x_2 \). (b) Pseudoconvexity and strict pseudoconvexity: \( f \) is pseudoconvex but not strictly pseudoconvex at \( x_1 \); \( f \) is both pseudoconvex and strictly pseudoconvex at \( x_2 \). (c) Quasiconvexity, strict quasiconvexity, and strong quasiconvexity: \( f \) is quasiconvex but neither strictly quasiconvex nor strongly quasiconvex at \( x_1 \); \( f \) is both quasiconvex and strictly quasiconvex at \( x_2 \) but not strongly quasiconvex at \( x_2 \); \( f \) is quasiconvex, strictly quasiconvex, and strongly quasiconvex at \( x_3 \).}
\end{figure}
We specify below some important results related to convexity of a function \( f \) at a point, where \( f : S \rightarrow \mathbb{R} \) and \( S \) is a nonempty convex set in \( \mathbb{R}^n \). Of course, not all the results developed throughout this chapter hold true. However, several of these results hold true and are summarized below. The proofs are similar to the corresponding theorems in this chapter.

1. Let \( f \) be both convex and differentiable at \( \bar{x} \). Then
   \[
   f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x})
   \]
   for each \( x \in S \). If \( f \) is strictly convex, strict inequality holds for \( x \neq \bar{x} \).
2. Let \( f \) be both convex and twice differentiable at \( \bar{x} \). Then the Hessian matrix \( H(\bar{x}) \) is positive semidefinite.
3. Let \( f \) be convex at \( \bar{x} \in S \), and let \( \bar{x} \) be an optimal solution to the problem to minimize \( f(x) \) subject to \( x \in S \). Then \( \bar{x} \) is a global optimal solution.
4. Let \( f \) be convex and differentiable at \( \bar{x} \in S \). Then \( \bar{x} \) is an optimal solution to the problem to minimize \( f(x) \) subject to \( x \in S \) if and only if \( \nabla f(\bar{x})^T (x - \bar{x}) \geq 0 \) for each \( x \in S \). In particular, if \( \bar{x} \in \text{int } S \), \( \bar{x} \) is an optimal solution if and only if \( \nabla f(\bar{x}) = 0 \).
5. Let \( f \) be convex and differentiable at \( \bar{x} \in S \). Suppose that \( \bar{x} \) is an optimal solution to the problem to maximize \( f(x) \) subject to \( x \in S \). Then \( \nabla f(\bar{x})^T (x - \bar{x}) \leq 0 \) for each \( x \in S \).
6. Let \( f \) be both quasiconvex and differentiable at \( \bar{x} \), and let \( x \in S \) be such that \( f(x) \leq f(\bar{x}) \). Then \( \nabla f(\bar{x})^T (x - \bar{x}) \leq 0 \).
7. Suppose that \( \bar{x} \) is a local optimal solution to the problem to minimize \( f(x) \) subject to \( x \in S \). If \( f \) is strictly quasiconvex at \( \bar{x} \), \( \bar{x} \) is a global optimal solution. If \( f \) is strongly quasiconvex at \( \bar{x} \), \( \bar{x} \) is the unique global optimal solution.
8. Consider the problem to minimize \( f(x) \) subject to \( x \in S \), and let \( \bar{x} \in S \) be such that \( \nabla f(\bar{x}) = 0 \). If \( f \) is pseudoconvex at \( \bar{x} \), \( \bar{x} \) is a global optimal solution; and if \( f \) is strictly pseudoconvex at \( \bar{x} \), \( \bar{x} \) is the unique global optimal solution.

**Exercises**

**[3.1]** Which of the following functions is convex, concave, or neither? Why?

a. \( f(x_1, x_2) = 2x_1^2 - 4x_1x_2 - 8x_1 + 3x_2 \)
b. \( f(x_1, x_2) = x_1 e^{-(x_1 + 3x_2)} \)
c. \( f(x_1, x_2) = -x_1^2 - 3x_2^2 + 4x_1x_2 + 10x_1 - 10x_2 \)
d. \( f(x_1, x_2, x_3) = 2x_1x_2 + 2x_1^2 + x_2^2 + 2x_3^2 - 5x_1x_3 \)
Chapter 3

3.2 Over what subset of \( \{ x : x > 0 \} \) is the univariate function \( f(x) = e^{-ax^b} \) convex, where \( a > 0 \) and \( b \geq 1 \)?

3.3 Prove or disprove concavity of the following function defined over \( S = \{(x_1, x_2) : -1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1\} \):

\[
f(x_1, x_2) = 10 - 3(x_2 - x_1^2)^2.
\]

Repeat for a convex set \( S \subseteq \{(x_1, x_2) : x_1^2 \geq x_2\} \).

3.4 Over what domain is the function \( f(x) = x^2(x^2 - 1) \) convex? Is it strictly convex over the region(s) specified? Justify your answer.

3.5 Show that a function \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) is affine if and only if \( f \) is both convex and concave. [A function \( f \) is affine if it is of the form \( f(x) = \alpha + \mathbf{c}'x \), where \( \alpha \) is a scalar and \( \mathbf{c} \) is an \( n \)-vector.]

3.6 Let \( S \) be a nonempty convex set in \( \mathbb{R}^n \), and let \( f: S \rightarrow \mathbb{R} \). Show that \( f \) is convex if and only if for any integer \( k \geq 2 \), the following holds true: \( x_1, \ldots, x_k \in S \) implies that \( f(\sum_{j=1}^{k} \lambda_j x_j) \leq \sum_{j=1}^{k} \lambda_j f(x_j) \), where \( \sum_{j=1}^{k} \lambda_j = 1, \lambda_j \geq 0 \) for \( j = 1, \ldots, k \).

3.7 Let \( S \) be a nonempty convex set in \( \mathbb{R}^n \), and let \( f: S \rightarrow \mathbb{R} \). Show that \( f \) is concave if and only if \( \text{hyp} \) \( f \) is convex.

3.8 Let \( f_1, f_2, \ldots, f_k: \mathbb{R}^n \rightarrow \mathbb{R} \) be convex functions. Consider the function \( f \) defined by \( f(x) = \sum_{j=1}^{k} \alpha_j f_j(x) \), where \( \alpha_j > 0 \) for \( j = 1, 2, \ldots, k \). Show that \( f \) is convex. State and prove a similar result for concave functions.

3.9 Let \( f_1, f_2, \ldots, f_k: \mathbb{R}^n \rightarrow \mathbb{R} \) be convex functions. Consider the function \( f \) defined by \( f(x) = \max\{f_1(x), f_2(x), \ldots, f_k(x)\} \). Show that \( f \) is convex. State and prove a similar result for concave functions.

3.10 Let \( h: \mathbb{R}^n \rightarrow \mathbb{R} \) be a convex function, and let \( g: \mathbb{R} \rightarrow \mathbb{R} \) be a nondecreasing convex function. Consider the composite function \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) defined by \( f(x) = g[h(x)] \). Show that \( f \) is convex.

3.11 Let \( g: \mathbb{R}^n \rightarrow \mathbb{R} \) be a concave function, and let \( f \) be defined by \( f(x) = 1/g(x) \). Show that \( f \) is convex over \( S = \{x : g(x) > 0\} \). State a symmetric result interchanging the convex and concave functions.

3.12 Let \( S \) be a nonempty convex set in \( \mathbb{R}^n \), and let \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) be defined as follows:

\[
f(x_1, x_2, x_3) = -2x_1^2 - 3x_2^2 - 2x_3^2 + 8x_1x_2 + 3x_1x_3 + 4x_2x_3
\]
\[ f(y) = \inf\{\|y - x\| : x \in S\}. \]

Note that \( f(y) \) gives the distance from \( y \) to the set \( S \) and is called the distance function. Prove that \( f \) is convex.

[3.13] Let \( S = \{(x_1, x_2) : x_1^2 + x_2^2 \leq 4\} \). Let \( f \) be the distance function defined in Exercise 3.12. Find the function \( f \) explicitly.

[3.14] Let \( S \) be a nonempty, bounded convex set in \( \mathbb{R}^n \), and let \( f: \mathbb{R}^n \to \mathbb{R} \) be defined as follows:

\[ f(y) = \sup\{y^t x : x \in S\}. \]

The function \( f \) is called the support function of \( S \). Prove that \( f \) is convex. Also, show that if \( f(y) = y^t \bar{x} \), where \( \bar{x} \in S \), \( \bar{x} \) is a subgradient of \( f \) at \( y \).

[3.15] Let \( S = A \cup B \), where

\[ A = \{(x_1, x_2) : x_1 < 0, x_1^2 + x_2^2 \leq 4\} \]
\[ B = \{(x_1, x_2) : x_1 \geq 0, -2 \leq x_2 \leq 2\}. \]

Find the support function defined in Exercise 3.14 explicitly.

[3.16] Let \( g: \mathbb{R}^m \to \mathbb{R} \) be a convex function, and let \( h: \mathbb{R}^n \to \mathbb{R}^m \) be an affine function of the form \( h(x) = Ax + b \), where \( A \) is an \( m \times n \) matrix and \( b \) is an \( m \times 1 \) vector. Then show that the composite function \( f: \mathbb{R}^n \to \mathbb{R} \) defined as \( f(x) = g[h(x)] \) is a convex function. Also, assuming twice differentiability of \( g \), derive an expression for the Hessian of \( f \).

[3.17] Let \( F \) be a cumulative distribution function for a random variable \( b \), that is, \( F(y) = \text{Prob}(b \leq y) \). Show that \( \phi(z) = \int_z^\infty F(y) \, dy \) is a convex function. Is \( \phi \) convex for any nondecreasing function \( F \)?

[3.18] A function \( f: \mathbb{R}^n \to \mathbb{R} \) is called a gauge function if it satisfies the following equality:

\[ f(\lambda x) = \lambda f(x) \quad \text{for all} \ x \in \mathbb{R}^n \text{ and all} \ \lambda \geq 0. \]

Further, a gauge function is said to be subadditive if it satisfies the following inequality:

\[ f(x) + f(y) \geq f(x + y) \quad \text{for all} \ x, y \in \mathbb{R}^n. \]

Prove that subadditivity is equivalent to convexity of gauge functions.

[3.19] Let \( f: S \to \mathbb{R} \) be defined as
where $S$ is a convex subset of $\mathbb{R}^n$, $\alpha$ and $\beta$ are vectors in $\mathbb{R}^n$, and where $\beta^T x > 0$ for all $x \in S$. Derive an explicit expression for the Hessian of $f$, and hence verify that $f$ is convex over $S$.

**[3.20]** Consider a quadratic function $f: \mathbb{R}^n \to \mathbb{R}$ and suppose that $f$ is convex on $S$, where $S$ is a nonempty convex set in $\mathbb{R}^n$. Show that:

a. The function $f$ is convex on $M(S)$, where $M(S)$ is the affine manifold containing $S$ defined by $M(S) = \{y : y = \sum_{j=1}^{k} \lambda_j x_j, \sum_{j=1}^{k} \lambda_j = 1, x_j \in S$ for all $j$, for $k \geq 1\}$. 

b. The function $f$ is convex on $L(S)$, the linear subspace parallel to $M(S)$, defined by $L(S) = \{y - x : y \in M(S)$ and $x \in S\}$. (This result is credited to Cottle [1967].)

**[3.21]** Let $h: \mathbb{R}^n \to \mathbb{R}$ be convex, and let $A$ be an $m \times n$ matrix. Consider the function $h: \mathbb{R}^m \to \mathbb{R}$ defined as follows:

$$h(y) = \inf\{f(x) : Ax = y\}.$$ 

Show that $h$ is convex.

**[3.22]** Let $S$ be a nonempty convex set in $\mathbb{R}^n$, and let $f: \mathbb{R}^n \to \mathbb{R}$ and $g: \mathbb{R}^n \to \mathbb{R}$ be convex. Consider the perturbation function $\phi: \mathbb{R}^m \to \mathbb{R}$ defined below:

$$\phi(y) = \inf\{f(x) : g(x) \leq y, x \in S\}.$$

a. Prove that $\phi$ is convex.

b. Show that if $y_1 \leq y_2$, $\phi(y_1) \geq \phi(y_2)$.

**[3.23]** Let $f: \mathbb{R}^n \to \mathbb{R}$ be lower semicontinuous. Show that the level set $S_\alpha = \{x : f(x) \leq \alpha\}$ is closed for all $\alpha \in \mathbb{R}$.

**[3.24]** Let $f$ be a convex function on $\mathbb{R}^n$. Prove that the set of subgradients of $f$ at a given point forms a closed convex set.

**[3.25]** Let $f: \mathbb{R}^n \to \mathbb{R}$ be convex. Show that $\xi$ is a subgradient of $f$ at $\bar{x}$ if and only if the hyperplane $\{(x, y) : y = f(\bar{x}) + \xi^T (x - \bar{x})\}$ supports $\text{epi } f$ at $[\bar{x}, f(\bar{x})]$. State and prove a similar result for concave functions.

**[3.26]** Let $f: \mathbb{R}^n \to \mathbb{R}$ be defined by $f(x) = \|x\|$. Prove that subgradients of $f$ are characterized as follows: If $x = 0$, $\xi$ is a subgradient of $f$ at $x$ if and only if...
\[ \|\xi\| \leq 1. \text{ On the other hand, if } x \neq 0, \xi \text{ is a subgradient of } f \text{ at } x \text{ if and only if } \|\xi\| = 1 \text{ and } \xi^T x = \|x\|. \] Use this result to show that } f \text{ is differentiable at each } x \neq 0, \text{ and characterize the gradient vector.}

[3.27] Let } f_1, f_2 : \mathbb{R}^n \rightarrow \mathbb{R} \text{ be differentiable convex functions. Consider the function } f \text{ defined by } f(x) = \max\{f_1(x), f_2(x)\}. \text{ Let } \bar{x} \text{ be such that } f(\bar{x}) = f_1(\bar{x}) = f_2(\bar{x}). \text{ Show that } \xi \text{ is a subgradient of } f \text{ at } \bar{x} \text{ if and only if }
\[ \xi = \lambda \nabla f_1(\bar{x}) + (1 - \lambda) \nabla f_2(\bar{x}), \text{ where } \lambda \in [0, 1]. \]
Generalize the result to several convex functions and state a similar result for concave functions.

[3.28] Consider the function } \theta \text{ defined by the following optimization problem for any } u \geq 0, \text{ where } X \text{ is a compact polyhedral set.}
\[ \theta(u) = \text{Minimize } c^T x + u^T (Ax - b) \]
subject to } x \in X.

a. Show that } \theta \text{ is concave.

b. Characterize the subgradients of } \theta \text{ at any given } u.

[3.29] In reference to Exercise 3.28, find the function } \theta \text{ explicitly and describe the set of subgradients at each point } u \geq 0 \text{ if }
\[ A = \begin{bmatrix} 3 & 2 \\ -1 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 6 \\ 4 \end{bmatrix}, \quad c = \begin{bmatrix} -1 \\ -2 \end{bmatrix} \]
\[ X = \{(x_1, x_2) : 0 \leq x_1 \leq 3/2, \ 0 \leq x_2 \leq 3/2\}. \]

[3.30] Consider the function } \theta \text{ defined by the following optimization problem:
\[ \theta(u_1, u_2) = \text{Minimize } x_1(2 - u_1) + x_2(3 - u_2) \]
subject to } x_1^2 + x_2^2 \leq 4.

a. Show that } \theta \text{ is concave.

b. Evaluate } \theta \text{ at the point } (2, 3).

c. Find the collection of subgradients of } \theta \text{ at } (2, 3).

[3.31] Let } f : S \rightarrow \mathbb{R}, \text{ where } S \subseteq \mathbb{R}^n \text{ is a nonempty convex set. Then the convex envelope of } f \text{ over } S, \text{ denoted } f_S(x), x \in S, \text{ is a convex function such that } f_S(x) \leq f(x) \text{ for all } x \in S; \text{ and if } g \text{ is any other convex function for which } g(x) \leq f(x) \text{ for all } x \in S, f_S(x) \geq g(x) \text{ for all } x \in S. \text{ Hence } f_S \text{ is the pointwise supremum over all convex underestimators of } f \text{ over } S. \text{ Show that } \min\{f(x) : x \in S\} = \min\{f_S(x) : x \in S\}, \text{ assuming that the minima exist, and that
\{x^* \in S : f(x^*) \leq f(x) \text{ for all } x \in S \}
\subseteq \{x^* \in S : f_S(x^*) \leq f_S(x) \text{ for all } x \in S \}.

**[3.32]** Let \( f : S \to R \) be a concave function, where \( S \subseteq R^n \) is a nonempty polytope with vertices \( x_1, \ldots, x_E \). Show that the convex envelope (see Exercise 3.31) of \( f \) over \( S \) is given by

\[
f_S(x) = \min \left\{ \sum_{i=1}^E \lambda_i f(x_i) : \sum_{i=1}^E \lambda_i x_i = x, \sum_{i=1}^E \lambda_i = 1, \lambda_i \geq 0 \text{ for } i = 1, \ldots, E \right\}.
\]

Hence, show that if \( S \) is a simplex in \( R^n \), \( f_S \) is an affine function that attains the same values as \( f \) over all the vertices of \( S \). (This result is due to Falk and Hoffman [1976].)

**[3.33]** Let \( f : S \to R \) and \( f_S : S \to R \) be as defined in Exercise 3.31. Show that if \( f \) is continuous, the epigraph \( \{(x,y) : y \geq f_S(x), \ x \in S, \ y \in R \} \) of \( f_S \) over \( S \) is the closure of the convex hull of the epigraph \( \{(x,y) : y \geq f(x), \ x \in S, \ y \in R \} \) of \( f \) over \( S \). Give an example to show that the epigraph of the latter set is not necessarily closed.

**[3.34]** Let \( f(x,y) = xy \) be a bivariate bilinear function, and let \( S \) be a polytope in \( R^2 \) having no edge with a finite, positive slope. Define \( \Lambda = \{ (\alpha, \beta, \gamma) \in R^3 : \alpha x_k + \beta y_k + \gamma \leq x_k y_k \text{ for } k = 1, \ldots, K \} \), where \( (x_k, y_k), \ k = 1, \ldots, K \), are the vertices of \( S \). Referring to Exercise 3.31, show that if \( S \) is two-dimensional, the set of extreme points \( (\alpha_e, \beta_e, \gamma_e), \ e = 1, \ldots, E \), of \( \Lambda \) is nonempty and that

\[
f_S(x,y) = \max \{ \alpha_e x + \beta_e y + \gamma_e, e = 1, \ldots, E \}.
\]

On the other hand, if \( S \) is one-dimensional and given by the convex hull of \( (x_1, y_1) \) and \( (x_2, y_2) \), show that there exists a solution \( (\alpha_1, \beta_1, \gamma_1) \) to the system \( \alpha x_k + \beta y_k + \gamma = x_k y_k \) for \( k = 1, 2 \), and in this case, \( f_S(x,y) = \alpha_1 x + \beta_1 y + \gamma_1 \). Specialize this result to verify that if \( S = \{(x,y) : a \leq x \leq b, \ c \leq y \leq d \} \), where \( a < b \) and \( c < d \), then

\[
f_S(x,y) = \max \{ ax + by + bd, \ cx + ay + ac \}.
\]

(This result is due to Sherali and Alameddine [1990].)

**[3.35]** Consider a triangle \( S \) having vertices \( (0,1) \), \( (2,0) \), and \( (1,2) \) and let \( f(x,y) = xy \) be a bivariate, bilinear function. Show that the convex envelope \( f_S \) of \( f \) over \( S \) (see Exercise 3.31) is given by

\[
f_S(x,y) = \begin{cases} 
-y + \frac{3y^2}{2 - x + y} & \text{for } (x,y) \neq (2,0) \\
0 & \text{for } (x,y) = (2,0)
\end{cases}
\]
Can you generalize your approach to finding the convex envelope of \( f \) over a triangle having a single edge that has a finite, positive slope? (This result is due to Sherali and Alameddine [1990].)

[3.36] Let \( f: \mathbb{R}^n \to \mathbb{R} \) be a differentiable function. Show that the gradient vector is given by

\[
\nabla f(x) = \left( \frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n} \right)'.
\]

[3.37] Let \( f: \mathbb{R}^n \to \mathbb{R} \) be a differentiable function. The linear approximation of \( f \) at a given point \( \bar{x} \) is given by

\[
 f(\bar{x}) + \nabla f(\bar{x})'(x - \bar{x}).
\]

If \( f \) is twice differentiable at \( \bar{x} \), the quadratic approximation of \( f \) at \( \bar{x} \) is given by

\[
 f(\bar{x}) + \nabla f(\bar{x})'(x - \bar{x}) + \frac{1}{2}(x - \bar{x})'H(\bar{x})(x - \bar{x}).
\]

Let \( f(x_1, x_2) = e^{2x_1^2-x_2^2} - 3x_1 + 5x_2 \). Give the linear and quadratic approximations of \( f \) at \((1, 1)\). Are these approximations convex, concave, or neither? Why?

[3.38] Consider the function \( f: \mathbb{R}^n \to \mathbb{R} \), and suppose that \( f \) is infinitely differentiable. Then show that \( f \) is strictly convex if and only if for each \( \bar{x} \) and \( d \) in \( \mathbb{R}^n \), the first nonzero derivative term of order greater than or equal to 2 in the Taylor series expansion exists, is of even order, and is positive.

[3.39] Consider the function \( f: \mathbb{R}^3 \to \mathbb{R} \), given by \( f(x) = x'Ax \), where

\[
 A = \begin{bmatrix} 2 & 2 & 3 \\ 1 & 3 & 1 \\ 1 & 2 & \theta \end{bmatrix}.
\]

What is the Hessian of \( f \)? For what values of \( \theta \) is \( f \) strictly convex?

[3.40] Consider the function \( f(x) = x^3 \), defined over the set \( S = \{ x \in \mathbb{R} : x \geq 0 \} \). Show that \( f \) is strictly convex over \( S \). Noting that \( f''(0) = 0 \) and \( f'''(0) = 6 \), comment on the application of Theorem 3.3.9.

[3.41] Let \( H \) be an \( n \times n \) symmetric, positive semidefinite matrix, and suppose that \( x'Hx = 0 \) for some \( x \in \mathbb{R}^n \). Then show that \( Hx = 0 \). (Hint: Consider the diagonal of the quadratic form \( x'Hx \) via the transformation \( x = Qy \), where the columns of \( Q \) are the normalized eigenvectors of \( H \).)
Let $H$ be an $n \times n$ symmetric matrix. Using the eigenvalue characterization of definiteness, verify that $H$ is positive definite if and only if it is positive semidefinite and nonsingular.

Suppose that $H$ is an $n \times n$ symmetric matrix. Show how Theorem 3.3.12 demonstrates that $H$ is positive definite if and only if it can be premultiplied by a series of $n$ lower triangular Gauss–Jordan reduction matrices $L_1, \ldots, L_n$ to yield an upper triangular matrix $U$ with positive diagonal elements. (Letting $L^{-1} = L_n^{-1} \cdots L_1^{-1}$, we obtain $H = LU$, where $L$ is lower triangular. This is known as the LU-decomposition of $H$; see Appendix A.2.) Furthermore, show that $H$ is positive definite if and only if there exists a lower triangular matrix $L$ with positive diagonal elements such that $H = LL^T$. (This is known as the Cholesky factorization of $H$; see Appendix A.2.)

Suppose that $S \neq \emptyset$ is closed and convex. Let $f$ be differentiable on int $S$. State if the following are true or false, justifying your answer:

a. If $f$ is convex on $S$, $f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x})$ for all $x \in S$, $\bar{x} \in \text{int } S$.

b. If $f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x})$ for all $x \in S$ and $\bar{x} \in \text{int } S$, $f$ is convex on $S$.

Consider the following problem:

\[
\text{Minimize } (x_1 - 4)^2 + (x_2 - 6)^2 \\
\text{subject to } \begin{cases} 
 x_2 \geq x_1^2 \\
 x_2 \leq 4.
\end{cases}
\]

Write a necessary condition for optimality and verify that it is satisfied by the point $(2, 4)$. Is this the optimal point? Why?

Use Theorem 3.4.3 to prove that every local minimum of a convex function over a convex set is also a global minimum.

Consider the problem to minimize $\{f(x) : x \in S\}$ and suppose that there exists an $\varepsilon > 0$ such that $N_\varepsilon(\bar{x}) \cap S$ is a convex set and that $f(\bar{x}) \leq f(x)$ for all $x \in N_\varepsilon(\bar{x}) \cap S$.

a. Show that if $H(\bar{x})$ is positive definite, $\bar{x}$ is both a strict and a strong local minimum.

b. Show that if $\bar{x}$ is a strict local minimum and $f$ is pseudoconvex on $N_\varepsilon(\bar{x}) \cap S$, $\bar{x}$ is also a strong local minimum.

Let $f: \mathbb{R}^n \to \mathbb{R}$ be a convex function, and suppose that $f(x + \lambda d) \geq f(x)$ for all $\lambda \in (0, \delta)$, where $\delta > 0$. Show that $f(x + \lambda d)$ is a nondecreasing function of $\lambda$. In particular, show that $f(x + \lambda d)$ is a strictly increasing function of $\lambda$ if $f$ is strictly convex.

Consider the following problem:
Maximize \( c'x + \frac{1}{2}x'Hx \)
subject to \( Ax \leq b \)
\( x \geq 0, \)

where \( H \) is a symmetric negative definite matrix, \( A \) is an \( m \times n \) matrix, \( c \) is an \( n \)-vector, and \( b \) is an \( m \)-vector. Write the necessary and sufficient condition for optimality of Theorem 3.4.3, and simplify it using the special structure of this problem.

Consider the problem to minimize \( f(x) \) subject to \( x \in S \), where \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) is a differentiable convex function and \( S \) is a nonempty convex set in \( \mathbb{R}^n \). Prove that \( \bar{x} \) is an optimal solution if and only if \( \nabla f(\bar{x})'(x - \bar{x}) \geq 0 \) for each \( x \in S \). State and prove a similar result for the maximization of a concave function. (This result was proved in the text as Corollary 2 to Theorem 3.4.3. In this exercise the reader is asked to give a direct proof without resorting to subgradients.)

A vector \( d \) is called a direction of descent of \( f \) at \( \bar{x} \) if there exists a \( \delta > 0 \) such that \( f(\bar{x} + \lambda d) < f(\bar{x}) \) for each \( \lambda \in (0, \delta) \). Suppose that \( f \) is convex. Show that \( d \) is a direction of descent if and only if \( f'(\bar{x};d) < 0 \). Does the result hold true without the convexity of \( f \)?

Consider the following problem:

Maximize \( f(x) \)
subject to \( Ax = b \)
\( x \geq 0, \)

where \( A \) is an \( m \times n \) matrix with rank \( m \) and \( f \) is a differentiable convex function. Consider the extreme point \( (x_B', x_N') = (\bar{b}' , 0') \), where \( \bar{b} = B^{-1}b \geq 0 \) and \( A = [B, N] \). Decompose \( \nabla f(x) \) accordingly into \( \nabla_B f(x) \) and \( \nabla_N f(x) \). Show that the necessary condition of Theorem 3.4.6 holds true if \( \nabla_N f(x)' - \nabla_B f(x)'B^{-1}N \leq 0 \). If this condition holds, is it necessarily true that \( x \) is a local maximum? Prove or give a counterexample.

If \( \nabla_N f(x)' - \nabla_B f(x)'B^{-1}N \leq 0 \), choose a positive component \( j \) and increase its corresponding nonbasic variable \( x_j \) until a new extreme point is reached. Show that this process results in a new extreme point having a larger objective value. Does this method guarantee convergence to a global optimal solution? Prove or give a counterexample.

Apply the procedure of Exercise 3.52 to the following problem starting with the extreme point \((1/2, 3, 0, 0)\):
Maximize \((x_1 - 2)^2 + (x_2 - 5)^2\)
subject to \(-2x_1 + x_2 + x_3 = 2\)
\(2x_1 + 3x_2 + x_4 = 10\)
\(x_1, x_2, x_3, x_4 \geq 0.\)

\[\text{Chapter 3}\]

[3.54] Consider the problem to minimize \(f(x)\) subject to \(x \in S\), where \(f: R^n \rightarrow R\) is convex and \(S\) is a nonempty convex set in \(R^n\). The cone of feasible directions of \(S\) at \(x \in S\) is defined by

\[D = \{d : \text{there exists a } \delta > 0 \text{ such that } \bar{x} + \lambda d \in S \text{ for } \lambda \in (0, \delta)\}.\]

Show that \(\bar{x}\) is an optimal solution to the problem if and only if \(f'(S; d) \geq 0\) for each \(d \in D\). Compare this result with the necessary and sufficient condition of Theorem 3.4.3. Specialize the result to the case where \(S = R^n\).

[3.55] Let \(f: R^n \rightarrow R\) be a quadratic function. Show that \(f\) is quasiconvex if and only if it is strictly quasiconvex, which holds true if and only if it is pseudoconvex. Furthermore, show that \(f\) is strongly quasiconvex if and only if it is strictly pseudoconvex.

[3.56] Let \(h: R^n \rightarrow R\) be a quasiconvex function, and let \(g: R \rightarrow R\) be a nondecreasing function. Then show that the composite function \(f: R^n \rightarrow R\) defined as \(f(x) = g[h(x)]\) is quasiconvex.

[3.57] Let \(f: S \subseteq R \rightarrow R\) be a univariate function, where \(S\) is some interval on the real line. Define \(f\) as \textit{unimodal} on \(S\) if there exists an \(x^* \in S\) at which \(f\) attains a minimum and \(f\) is nondecreasing on the interval \(\{x \in S : x \geq x^*\}\), whereas it is nonincreasing on the interval \(\{x \in S : x < x^*\}\). Assuming that \(f\) attains a minimum on \(S\), show that \(f\) is quasiconvex if and only if it is unimodal on \(S\).

[3.58] Let \(f: S \rightarrow R\) be a continuous function, where \(S\) is a convex subset of \(R^n\). Show that \(f\) is quasimonotone if and only if the level surface \(\{x \in S : f(x) = \alpha\}\) is a convex set for all \(\alpha \in R\).

[3.59] Let \(f: S \rightarrow R\) be a differentiable function, where \(S\) is an open, convex subset of \(R^n\). Show that \(f\) is quasimonotone if and only if for every \(x_1\) and \(x_2\) in \(S\), \(f(x_1) \geq f(x_2)\) implies that \(\nabla f(x_2)'(x_1 - x_2) \geq 0\) and \(f(x_1) \leq f(x_2)\) implies that \(\nabla f(x_2)'(x_1 - x_2) \leq 0\). Hence, show that \(f\) is quasimonotone if and only if \(f(x_1) \geq f(x_2)\) implies that \(\nabla f(x_2)'(x_1 - x_2) \geq 0\) for all \(x_1\) and \(x_2\) in \(S\) and for all \(x_\lambda = \lambda x_1 + (1-\lambda)x_2\), where \(0 \leq \lambda \leq 1\).

[3.60] Let \(f: S \rightarrow R\), where \(f\) is lower semicontinuous and where \(S\) is a convex subset of \(R^n\). Define \(f\) as being \textit{strongly unimodal} on \(S\) if for each \(x_1\) and \(x_2\) in
S for which the function $F(\lambda) = f[x_1 + \lambda(x_2 - x_1)]$, $0 \leq \lambda \leq 1$, attains a minimum at a point $\lambda^* > 0$, we have that $F(0) > F(\lambda) > F(\lambda^*)$ for all $0 < \lambda < \lambda^*$. Show that $f$ is strongly quasiconvex on $S$ if and only if it is strongly unimodal on $S$ (see Exercise 8.10).

[3.61] Let $g: S \to R$ and $h: S \to R$, where $S$ is a nonempty convex set in $R^n$. Consider the function $f: S \to R$ defined by $f(x) = g(x)/h(x)$. Show that $f$ is quasiconvex if the following two conditions hold true:

a. $g$ is convex on $S$, and $g(x) \geq 0$ for each $x \in S$.

b. $h$ is concave on $S$, and $h(x) > 0$ for each $x \in S$.

(Hint: Use Theorem 3.5.2.)

[3.62] Show that the function $f$ defined in Exercise 3.61 is quasiconvex if the following two conditions hold true:

a. $g$ is convex on $S$, and $g(x) \leq 0$ for each $x \in S$.

b. $h$ is convex on $S$, and $h(x) > 0$ for each $x \in S$.

[3.63] Let $g: S \to R$ and $h: S \to R$, where $S$ is a nonempty convex set in $R^n$. Consider the function $f: S \to R$ defined by $f(x) = g(x)/h(x)$. Show that $f$ is quasiconvex if the following two conditions hold true:

a. $g$ is convex, and $g(x) \leq 0$ for each $x \in S$.

b. $h$ is concave, and $h(x) > 0$ for each $x \in S$.

[3.64] In each of Exercises 3.61, 3.62, and 3.63, show that $f$ is pseudoconvex provided that $S$ is open and that $g$ and $h$ are differentiable.

[3.65] Let $e_1, e_2$ be nonzero vectors in $R^n$, and $\alpha_1, \alpha_2$ be scalars. Let $S = \{x: e_2^Tx + \alpha_2 > 0\}$. Consider the function $f: S \to R$ defined as follows:

$$f(x) = \frac{e_1^Tx + \alpha_1}{e_2^Tx + \alpha_2}.$$  

Show that $f$ is both pseudoconvex and pseudoconcave. (Functions that are both pseudoconvex and pseudoconcave are called pseudolinear.)

[3.66] Consider the quadratic function $f: R^n \to R$ defined by $f(x) = x^THx$. The function $f$ is said to be positive subdefinite if $x^THx < 0$ implies that $Hx \geq 0$ or $Hx \leq 0$ for each $x \in R^n$. Prove that $f$ is quasiconvex on the nonnegative orthant, $R^n_+ = \{x \in R^n : x \geq 0\}$ if and only if it is positive subdefinite. (This result is credited to Martos [1969].)

[3.67] The function $f$ defined in Exercise 3.66 is said to be strictly positive subdefinite if $x^THx < 0$ implies that $Hx > 0$ or $Hx < 0$ for each $x \in R^n$. Prove that $f$ is pseudoconvex on the nonnegative orthant excluding $x = 0$ if and only if it is strictly positive subdefinite. (This result is credited to Martos [1969].)
13.68. Let $S \rightarrow R$ be a continuously differentiable convex function, where $S$ is some open interval in $R$. Then show that $f$ is (strictly) pseudoconvex if and only if whenever $f'(x) = 0$ for any $x \in S$, this implies that $x$ is a (strict) local minimum of $f$ on $S$. Generalize this result to the multivariate case.

13.69. Let $S \rightarrow R$ be pseudoconvex, and suppose that for some $x_1$ and $x_2$ in $R^n$, we have $\nabla f(x_1)(x_2 - x_1) \geq 0$. Show that the function $F(\lambda) = f(x_1 + \lambda(x_2 - x_1))$ is nondecreasing for $\lambda \geq 0$.

13.70. Let $f: S \rightarrow R$ be a twice differentiable univariate function, where $S$ is some open interval in $R$. Then show that $f$ is (strictly) pseudoconvex if and only if whenever $f''(\bar{x}) = 0$ for any $\bar{x} \in S$, we have that either $f''(\bar{x}) > 0$ or that $f''(\bar{x}) = 0$ and $\bar{x}$ is a (strict) local minimum of $f$ over $S$. Generalize this result to the multivariate case.

13.71. Let $g_1, g_2: R^n \rightarrow R$ be differentiable and convex. Let $\alpha \in [0,1]$. Consider the function $G_{\alpha}: R^n \rightarrow R$ defined as

$$G_{\alpha}(x) = \frac{1}{2} \left[ g_1(x) + g_2(x) - \sqrt{g_1^2(x) + g_2^2(x) - 2\alpha g_1(x)g_2(x)} \right],$$

where $\sqrt{\cdot}$ denotes the positive square root.

a. Show that $G_{\alpha}(x) \geq 0$ if and only if $g_1(x) \geq 0$ and $g_2(x) \geq 0$, that is, minimum $\{g_1(x), g_2(x)\} \geq 0$.

b. If $g_1$ and $g_2$ are differentiable, show that $G_{\alpha}$ is differentiable at $x$ for each $\alpha \in [0,1]$ provided that $g_1(x), g_2(x) \neq 0$.

c. Now suppose that $g_1$ and $g_2$ are concave. Show that $G_{\alpha}$ is concave for $\alpha \in [0,1]$. Does this result hold true for $\alpha \in (-1,0)$?

d. Suppose that $g_1$ and $g_2$ are quasiconcave. Show that $G_{\alpha}$ is quasiconcave for $\alpha = 1$.

e. Let $g_1(x) = -x_1^2 - x_2^2 + 4$ and $g_2(x) = 2x_1 + x_2 - 1$. Obtain an explicit expression for $G_{\alpha}$, and verify parts a, b, and c.

This exercise describes a general method for combining two constraints of the form $g_1(x) \geq 0$ and $g_2(x) \geq 0$ into an equivalent single constraint of the
Convex Functions and Generalizations

form \( G_\alpha(x) \geq 0 \). This procedure could be applied successively to reduce a problem with several constraints into an equivalent single constrained problem. The procedure is due to Rvačev [1963].

[3.73] Let \( g_1, g_2 : \mathbb{R}^n \to \mathbb{R} \), and let \( \alpha \in [0,1] \). Consider the function \( G_\alpha : \mathbb{R}^n \to \mathbb{R} \) defined by

\[
G_\alpha(x) = \frac{1}{2} \left[ g_1(x) + g_2(x) + \sqrt{g_1^2(x) + g_2^2(x) - 2\alpha g_1(x)g_2(x)} \right],
\]

where \( \sqrt{\cdot} \) denotes the positive square root.

a. Show that \( G_\alpha(x) \geq 0 \) if and only if maximum \( \{g_1(x), g_2(x)\} \geq 0 \).

b. If \( g_1 \) and \( g_2 \) are differentiable, show that \( G_\alpha \) is differentiable at \( x \) for each \( \alpha \in [0,1] \), provided that \( g_1(x), g_2(x) \neq 0 \).

c. Now suppose that \( g_1 \) and \( g_2 \) are convex. Show that \( G_\alpha \) is convex for \( \alpha \in [0,1] \). Does the result hold true for \( \alpha \in (-1,0) \)?

d. Suppose that \( g_1 \) and \( g_2 \) are quasiconvex. Show that \( G_\alpha \) is quasiconvex for \( \alpha = 1 \).

e. In some optimization problems, the restriction that the variable \( x = 0 \) or 1 arises. Show that this restriction is equivalent to maximum \( \{g_1(x), g_2(x)\} \geq 0 \), where \( g_1(x) = -x^2 \) and \( g_2(x) = -(x-1)^2 \).

Find the function \( G_\alpha \) explicitly, and verify statements a, b, and c.

This exercise describes a general method for combining the either-or constraints of the form \( g_1(x) \geq 0 \) or \( g_2(x) \geq 0 \) into a single constraint of the form \( G_\alpha(x) \geq 0 \), and is due to Rvačev [1963].

Notes and References

In this chapter we deal with the important topic of convex and concave functions. The recognition of these functions is generally traced to Jensen [1905, 1906]. For earlier related works on the subject, see Hadamard [1893] and Hölder [1889].

In Section 3.1, several results related to continuity and directional derivatives of a convex function are presented. In particular, we show that a convex function is continuous on the interior of the domain. See, for example, Rockafellar [1970]. Rockafellar also discusses the convex extension to \( \mathbb{R}^n \) of a convex function \( f : S \subset \mathbb{R}^n \to \mathbb{R} \), which takes on finite values over a convex subset \( S \) of \( \mathbb{R}^n \), by letting \( f(x) = \infty \) for \( x \not\in S \). Accordingly, a set of arithmetic operations involving \( \infty \) also needs to be defined. In this case, \( S \) is referred to as the effective domain of \( f \). Also, a proper convex function is then defined as a convex function for which \( f(x) < \infty \) for at least one point \( x \) and for which \( f(x) > -\infty \) for all \( x \).
In Section 3.2 we discuss subgradients of convex functions. Many of the properties of differentiable convex functions are retained by replacing the gradient vector by a subgradient. For this reason, subgradients have been used frequently in the optimization of nondifferentiable functions. See, for example, Bertsekas [1975], Demyanov and Pallaschke [1985], Demyanov and Vasilev [1985], Held and Karp [1970], Held et al. [1974], Kiwiel [1985], Sherali et al. [2000], Shor [1985], and Wolfe [1976]. (See also, Chapter 8.)

In Section 3.3 we give some properties of differentiable convex functions. For further study of these topics as well as other properties of convex functions, refer to Eggleston [1958], Fenchel [1953], Roberts and Varberg [1973], and Rockafellar [1970]. The superdiagonalization algorithm derived from Theorem 3.3.12 provides an efficient polynomial-time algorithm for checking definiteness properties of matrices. This method is intimately related with LU and Cholesky factorization techniques (see Exercise 3.43, and refer to Section A.2, Fletcher [1985], Luenberger [1973a], and Murty [1983] for further details).

Section 3.4 treats the subject of minima and maxima of convex functions over convex sets. Robinson [1987] discusses the distinction between strict and strong local minima. For general functions, the study of minima and maxima is quite complicated. As shown in Section 3.4, however, every local minimum of a convex function over a convex set is also a global minimum, and the maximum of a convex function over a convex set occurs at an extreme point. For an excellent study of optimization of convex functions, see Rockafellar [1970]. The characterization of the optimal solution set for convex programs is due to Mangasarian [1988]. This paper also extends the results given in Section 3.4 to subdifferentiable convex functions.

In Section 3.5 we examine other classes of functions that are related to convex functions; namely, quasiconvex and pseudoconvex functions. The class of quasiconvex functions was first studied by De Finetti [1949]. Arrow and Enthoven [1961] derived necessary and sufficient conditions for quasiconvexity on the nonnegative orthant assuming twice differentiability. Their results were extended by Ferland [1972]. Note that a local minimum of a quasiconvex function over a convex set is not necessarily a global minimum. This result holds true, however, for a strictly quasiconvex function. Ponstein [1967] introduced the concept of strongly quasiconvex functions, which ensures that the global minimum is unique, a property that is not enjoyed by strictly quasiconvex functions. The notion of pseudoconvexity was introduced by Mangasarian [1965]. The significance of the class of pseudoconvex functions stems from the fact that every point with a zero gradient is a global minimum. Matrix theoretic characterizations (see, e.g., Exercises 3.66 and 3.67) of quadratic pseudoconvex and quasiconvex functions have been presented by Cottle and Ferland [1972] and by Martos [1965, 1967b, 1969, 1975]. For further reading on this topic, refer to Avriel et al. [1988], Fenchel [1953], Greenberg and Pierskalla [1971], Karamardian [1967], Mangasarian [1969a], Ponstein [1967], Schaible [1981a,b], and Schaible and Ziemba [1981]. The last four references give excellent surveys on this topic, and the results of Exercises 3.55 to 3.60 and 3.68 to 3.70 are discussed in detail by Avriel et al. [1988] and Schaible [1981a,b]. Karamardian
and Schaible [1990] also present various tests for checking generalized properties for differentiable functions. See also Section B.2.

Exercises 3.31 to 3.34 deal with convex envelopes of nonconvex functions. This construct plays an important role in global optimization techniques for nonconvex programming problems. For additional information on this subject, we refer the reader to Al-Khayyal and Falk [1983], Falk [1976], Grotzinger [1985], Horst and Tuy [1990], Pardalos and Rosen [1987], Sherali [1997], and Sherali and Alameddine [1990].
Part 2

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

In Chapter 3 we derived an optimality condition for a problem of the following form: Minimize $f(x)$ subject to $x \in S$, where $f$ is a convex function and $S$ is a convex set. The necessary and sufficient condition for $x$ to solve the problem was shown to be

$$\nabla f(x)(x - \bar{x}) \geq 0 \quad \text{for all } x \in S.$$

In this chapter the nature of the set $S$ will be specified more explicitly in terms of inequality and/or equality constraints. A set of first-order necessary conditions are derived without any convexity assumptions that are sharper than the above in the sense that they explicitly consider the constraint functions and are more easily verifiable, since they deal with a system of equations. Under suitable convexity assumptions, these necessary conditions are also sufficient for optimality. These optimality conditions lead to classical or direct optimization techniques for solving unconstrained and constrained problems that construct these conditions and then attempt to find a solution to them. In contrast, we discuss several indirect methods in Chapters 8 through 11, which iteratively improve the current solution, converging to a point that can be shown to satisfy these optimality conditions. A discussion of second-order necessary and/or sufficient conditions for unconstrained as well as for constrained problems is also provided.

Readers who are unfamiliar with generalized convexity concepts from Section 3.5 may substitute any references to such properties by related convexity assumptions for ease in reading.

Following is an outline of the chapter.

Section 4.1: Unconstrained Problems We consider briefly optimality conditions for unconstrained problems. First- and second-order conditions are discussed.

Section 4.2: Problems Having Inequality Constraints Both the Fritz John (FJ) and the Karush–Kuhn–Tucker (KKT) conditions for problems having
inequality constraints are derived. The nature and value of solutions satisfying these conditions are emphasized.

Section 4.3: Problems Having Inequality and Equality Constraints  This section extends the results of the preceding section to problems having both inequality and equality constraints.

Section 4.4: Second-Order Necessary and Sufficient Optimality Conditions for Constrained Problems  Similar to the unconstrained case discussed in Section 4.1, we develop second-order necessary and sufficient optimality conditions as an extension to the first-order conditions developed in Sections 4.2 and 4.3 for inequality and equality constrained problems. Many results and algorithms in nonlinear programming assume the existence of a local optimal solution that satisfies the second-order sufficiency conditions.

4.1 Unconstrained Problems

An unconstrained problem is a problem of the form to minimize $f(x)$ without any constraints on the vector $x$. Unconstrained problems seldom arise in practical applications. However, we consider such problems here because optimality conditions for constrained problems become a logical extension of the conditions for unconstrained problems. Furthermore, as shown in Chapter 9, one strategy for solving a constrained problem is to solve a sequence of unconstrained problems. We recall below the definitions of local and global minima for unconstrained problems as a special case of Definition 3.4.1, where the set $S$ is replaced by $R^n$.

4.1.1 Definition

Consider the problem of minimizing $f(x)$ over $R^n$, and let $\bar{x} \in R^n$. If $f(\bar{x}) \leq f(x)$ for all $x \in R^n$, $\bar{x}$ is called a global minimum. If there exists an $\epsilon$-neighborhood $N_\epsilon(\bar{x})$ around $\bar{x}$ such that $f(\bar{x}) \leq f(x)$ for each $x \in N_\epsilon(\bar{x})$, $\bar{x}$ is called a local minimum, while if $f(\bar{x}) < f(x)$ for all $x \in N_\epsilon(\bar{x})$, $x \neq \bar{x}$, for some $\epsilon > 0$, $\bar{x}$ is called a strict local minimum. Clearly, a global minimum is also a local minimum.

Necessary Optimality Conditions

Given a point $\bar{x}$ in $R^n$, we wish to determine, if possible, whether or not the point is a local or a global minimum of a function $f$. For this purpose we need to characterize a minimizing solution. Fortunately, the differentiability assumption of $f$ provides a means for obtaining this characterization. The corollary to Theorem 4.1.2 gives a first-order necessary condition for $\bar{x}$ to be a local optimum. Theorem 4.1.3 gives a second-order necessary condition using the Hessian matrix.
4.1.2 Theorem

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\overline{x}$. If there is a vector $d$ such that $\nabla f(\overline{x})'d < 0$, there exists a $\delta > 0$ such that $f(\overline{x} + \delta d) < f(\overline{x})$ for each $\delta \in (0, \delta)$, so that $d$ is a descent direction of $f$ at $\overline{x}$.

Proof

By the differentiability of $f$ at $\overline{x}$, we must have

$$f(\overline{x} + \lambda d) = f(\overline{x}) + \lambda \nabla f(\overline{x})'d + \lambda \|d\| \alpha(\overline{x}; \lambda d),$$

where $\alpha(\overline{x}; \lambda d) \to 0$ as $\lambda \to 0$. Rearranging the terms and dividing by $\lambda$, $\lambda \neq 0$, we get

$$\frac{f(\overline{x} + \lambda d) - f(\overline{x})}{\lambda} = \nabla f(\overline{x})'d + \|d\| \alpha(\overline{x}; \lambda d).$$

Since $\nabla f(\overline{x})'d < 0$ and $\alpha(\overline{x}; \lambda d) \to 0$ as $\lambda \to 0$, there exists a $\delta > 0$ such that $\nabla f(\overline{x})'d + \|d\| \alpha(\overline{x}; \lambda d) < 0$ for all $\lambda \in (0, \delta)$. The result then follows.

Corollary

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\overline{x}$. If $\overline{x}$ is a local minimum, $\nabla f(\overline{x}) = 0$.

Proof

Suppose that $\nabla f(\overline{x}) \neq 0$. Then, letting $d = -\nabla f(\overline{x})$, we get $\nabla f(\overline{x})'d = -\|\nabla f(\overline{x})\|^2 < 0$; and by Theorem 4.1.2, there is a $\delta > 0$ such that $f(\overline{x} + \lambda d) < f(\overline{x})$ for $\lambda \in (0, \delta)$, contradicting the assumption that $\overline{x}$ is a local minimum. Hence, $\nabla f(\overline{x}) = 0$.

The condition above uses the gradient vector whose components are the first partials of $f$. Hence, it is called a first-order condition. Necessary conditions can also be stated in terms of the Hessian matrix $H$, whose elements are the second partials of $f$, and are then called second-order conditions. One such condition is given below.

4.1.3 Theorem

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable at $\overline{x}$. If $\overline{x}$ is a local minimum, $\nabla f(\overline{x}) = 0$ and $H(\overline{x})$ is positive semidefinite.
**Proof**

Consider an arbitrary direction \( d \). Then from the differentiability of \( f \) at \( \bar{x} \), we have

\[
f(\bar{x} + \lambda d) = f(\bar{x}) + \lambda \nabla f(\bar{x})' d + \frac{1}{2} \lambda^2 d' H(\bar{x}) d + \lambda^2 \|d\|^2 \alpha(\bar{x}; \lambda d),
\]

where \( \alpha(\bar{x}; \lambda d) \to 0 \) as \( \lambda \to 0 \). Since \( \bar{x} \) is a local minimum, from the corollary to Theorem 4.1.2, we have \( \nabla f(\bar{x}) = 0 \). Rearranging the terms in (4.1) and dividing by \( \lambda^2 > 0 \), we get

\[
\frac{f(\bar{x} + \lambda d) - f(\bar{x})}{\lambda^2} = \frac{1}{2} d' H(\bar{x}) d + \|d\|^2 \alpha(\bar{x}; \lambda d).
\]

Since \( \bar{x} \) is a local minimum, \( f(\bar{x} + \lambda d) \geq f(\bar{x}) \) for \( \lambda \) sufficiently small. From (4.2) it is thus clear that \((1/2)d' H(\bar{x}) d + \|d\|^2 \alpha(\bar{x}; \lambda d) \geq 0 \) for \( \lambda \) sufficiently small. By taking the limit as \( \lambda \to 0 \), it follows that \( d' H(\bar{x}) d \geq 0 \); and hence, since \( d \) was arbitrary, \( H(\bar{x}) \) is positive semidefinite.

**Sufficient Optimality Conditions**

The conditions discussed thus far are necessary conditions; that is, they must be true for every local optimal solution. On the other hand, a point satisfying these conditions need not be a local minimum. Theorem 4.1.4 gives a sufficient condition for a local minimum.

**4.1.4 Theorem**

Suppose that \( f: \mathbb{R}^n \to \mathbb{R} \) is twice differentiable at \( \bar{x} \). If \( \nabla f(\bar{x}) = 0 \) and \( H(\bar{x}) \) is positive definite, \( \bar{x} \) is a strict local minimum.

**Proof**

Since \( f \) is twice differentiable at \( \bar{x} \), we must have, for each \( \bar{x} \in \mathbb{R}^n \),

\[
f(x) = f(\bar{x}) + \nabla f(\bar{x})'(x - \bar{x}) + \frac{1}{2} (x - \bar{x})' H(\bar{x})(x - \bar{x}) + \|x - \bar{x}\|^2 \alpha(\bar{x}; x - \bar{x}),
\]

where \( \alpha(\bar{x}; x - \bar{x}) \to 0 \) as \( x \to \bar{x} \). Suppose, by contradiction, that \( \bar{x} \) is not a strict local minimum; that is, suppose that there exists a sequence \( \{x_k\} \) converging to \( \bar{x} \) such that \( f(x_k) \leq f(\bar{x}) \), \( x_k \neq \bar{x} \), for each \( k \). Considering this sequence, noting that \( \nabla f(\bar{x}) = 0 \) and that \( f(x_k) \leq f(\bar{x}) \), and denoting \( (x_k - \bar{x})/\|x_k - \bar{x}\| \) by \( d_k \), (4.3) then implies that...
\frac{1}{2} \mathbf{d}_k^T \mathbf{H}(\bar{x}) \mathbf{d}_k + \alpha(\bar{x}; \mathbf{x}_k - \bar{x}) \leq 0 \quad \text{for each } k. \quad (4.4)

But \|\mathbf{d}_k\| = 1 for each \(k\); and hence there exists an index set \(\mathcal{I}\) such that \(\{\mathbf{d}_k\}_{\mathcal{I}}\) converges to \(\mathbf{d}\), where \(\|\mathbf{d}\| = 1\). Considering this subsequence and the fact that 
\[ \alpha(\bar{x}; \mathbf{x}_k - \bar{x}) \to 0 \quad \text{as } k \in \mathcal{I} \quad \text{approaches } \infty, \]
(4.4) implies that \(\mathbf{d}^T \mathbf{H}(\bar{x}) \mathbf{d} \leq 0\). This contradicts the assumption that \(\mathbf{H}(\bar{x})\) is positive definite since \(\|\mathbf{d}\| = 1\). Therefore, \(\bar{x}\) is indeed a strict local minimum.

Essentially, note that assuming \(f\) to be twice continuously differentiable, since \(\mathbf{H}(\bar{x})\) is positive definite, we have that \(\mathbf{H}(\mathbf{x})\) is positive definite in an \(\epsilon\)-neighborhood of \(\bar{x}\), so \(f\) is strictly convex in an \(\epsilon\)-neighborhood of \(\bar{x}\). Therefore, as follows from Theorem 3.4.2, \(\bar{x}\) is a strict local minimum, that is, it is the unique global minimum over \(N_\epsilon(\bar{x})\) for some \(\epsilon > 0\). In fact, noting the second part of Theorem 3.4.2, we can conclude that \(\bar{x}\) is also a strong or isolated local minimum in this case.

In Theorem 4.1.5, we show that the necessary condition \(\nabla f(\bar{x}) = 0\) is also sufficient for \(\bar{x}\) to be a global minimum if \(f\) is pseudoconvex at \(\bar{x}\). In particular, if \(\nabla f(\bar{x}) = 0\) and if \(\mathbf{H}(\mathbf{x})\) is positive semidefinite for all \(\mathbf{x}\), \(f\) is convex, and therefore also pseudoconvex. Consequently, \(\bar{x}\) is a global minimum. This is also evident from Theorem 3.3.3 or from Corollary 2 to Theorem 3.4.3.

**4.1.5 Theorem**

Let \(f \colon \mathbb{R}^n \to \mathbb{R}\) be pseudoconvex at \(\bar{x}\). Then \(\bar{x}\) is a global minimum if and only if \(\nabla f(\bar{x}) = 0\).

**Proof**

By the corollary to Theorem 4.1.2, if \(\bar{x}\) is a global minimum, \(\nabla f(\bar{x}) = 0\). Now suppose that \(\nabla f(\bar{x}) = 0\), so that \(\nabla f(\bar{x})^T (\mathbf{x} - \bar{x}) = 0\) for each \(\mathbf{x} \in \mathbb{R}^n\). By the pseudoconvexity of \(f\) at \(\bar{x}\), it then follows that \(f(\mathbf{x}) \geq f(\bar{x})\) for each \(\mathbf{x} \in \mathbb{R}^n\), and the proof is complete.

Theorem 4.1.5 provides a necessary and sufficient optimality condition in terms of the first-order derivative alone when \(f\) is pseudoconvex. In a similar manner, we can derive necessary and sufficient conditions for local optimality in terms of higher-order derivatives when \(f\) is infinitely differentiable, as an extension to the foregoing results. Toward this end, consider the following result for the univariate case.
4.1.6 Theorem

Let $f: \mathbb{R} \to \mathbb{R}$ be an infinitely differentiable univariate function. Then $x \in \mathbb{R}$ is a local minimum if and only if either $f^{(j)}(x) = 0$ for all $j = 1, 2, \ldots$, or else there exists an even $n \geq 2$ such that $f^{(n)}(x) > 0$ while $f^{(j)}(x) = 0$ for all $1 \leq j < n$, where $f^{(j)}$ denotes the $j$th-order derivative of $f$.

**Proof**

We know that $x$ is a local minimum of $f$ if and only if $f(x + h) - f(x) \geq 0$ for all sufficiently small values of $|h|$. Using the infinite Taylor series representation of $f(x + h)$, this holds true if and only if

$$hf^{(1)}(x) + \frac{h^2}{2!}f^{(2)}(x) + \frac{h^3}{3!}f^{(3)}(x) + \frac{h^4}{4!}f^{(4)}(x) + \cdots \geq 0$$

for all $|h|$ small enough. Similar to the proof of Theorem 3.3.9, it is readily verified that the foregoing inequality holds true if and only if the condition of the theorem is satisfied, and this completes the proof.

Before proceeding, we remark here that for a local maximum, the condition of Theorem 4.1.6 remains the same, except that we require $f^{(n)}(x) < 0$ in lieu of $f^{(n)}(x) > 0$. Observe also, noting Theorem 3.3.9, that the above result essentially asserts that for the case under discussion, $x$ is a local minimum if and only if $f$ is locally convex about $x$. This result can be partially extended, at least in theory, to the case of multivariate functions. Toward this end, suppose that $x \in \mathbb{R}^n$ is a local minimum for $f: \mathbb{R}^n \to \mathbb{R}$. Then this holds true if and only if $f(x + \lambda d) \geq f(x)$ for all $d \in \mathbb{R}^n$ and for all sufficiently small values of $|\lambda|$. Assuming $f$ to be infinitely differentiable, this asserts that for all $d \in \mathbb{R}^n$, $\|d\| = 1$, we must equivalently have

$$f(x + \lambda d) - f(x) = \lambda \nabla f(x)'d + \frac{\lambda^2}{2!}d' \mathbf{H}(x)d + \frac{\lambda^3}{3!} \sum_{i} \sum_{j} \sum_{k} f_{ijk}(x)d_id_jd_k + \cdots \geq 0$$

for all $-\delta \leq \lambda \leq \delta$, for some $\delta > 0$. Consequently, the first nonzero derivative term, if it exists, must correspond to an even power of $\lambda$ and must be positive in value.

Note that the foregoing concluding statement is not sufficient to claim local optimality of $x$. The difficulty is that it might be the case that this
The Fritz John and Karush–Kuhn–Tucker Optimality Conditions

statement holds true, implying that for any \( d \in \mathbb{R}^n \), \( \|d\| = 1 \), we have \( f(\bar{x} + \lambda d) \geq f(\bar{x}) \) for all \( -\delta_d \leq \lambda \leq \delta_d \) for some \( \delta_d > 0 \), which depends on \( d \), but then, \( \delta_d \) might get vanishingly small as \( d \) varies, so that we cannot assert the existence of a \( \delta > 0 \) such that \( f(\bar{x} + \lambda d) \geq f(\bar{x}) \) for all \( -\delta \leq \lambda \leq \delta \). In this case, by moving along curves instead of along straight lines, improving values of \( f \) might be accessible in the immediate neighborhood of \( \bar{x} \). On the other hand, a valid sufficient condition by Theorem 4.1.5 is that \( \nabla f(\bar{x}) = 0 \) and that \( f \) is convex (or pseudoconvex) over an \( \varepsilon \)-neighborhood about \( \bar{x} \), for some \( \varepsilon > 0 \). However, this might not be easy to check, and we might need to assess the situation numerically by examining values of the function at perturbations about the point \( \bar{x} \) (refer also to Exercise 4.19).

To illustrate the above point, consider the following example due to the Italian mathematician Peano. Let 
\[
 f(x_1, x_2) = (x_2 - x_1)(x_2^2 - 2x_1) = 2x_1^2 - 3x_1x_2 + x_2^4. 
\]
Then we have, at \( \bar{x} = (0, 0)' \),
\[
 \nabla f(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad H(0) = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix}, \quad f_{122}(0) = f_{212}(0) = f_{221}(0) = -6, \quad f_{2222}(0) = 24, 
\]
and all other partial derivatives of \( f \) of order 3 or higher are zeros. Hence, we obtain by the Taylor series expansion
\[
 f(\bar{x} + \lambda d) - f(\bar{x}) = \frac{\lambda^2}{2}(4d_1^2) + \frac{\lambda^3}{6}(-18d_1d_2^2) + \frac{\lambda^4}{24}(24d_2^4) 
= 2\lambda^2 \left( d_1 - \frac{3\lambda}{4} d_2 \right)^2 - \frac{1}{8} \lambda^4 d_2^4. 
\]
Note that for any \( d = (d_1, d_2)' \), \( \|d\| = 1 \), if \( d_1 \neq 0 \), the given necessary condition holds true because the second-order term is positive. On the other hand, if \( d_1 = 0 \), we must have \( d_2 \neq 0 \), and the condition holds true again because the first nonzero term is of order 4 and is positive. However, \( \bar{x} = (0, 0)' \) is not a local minimum, as evident from Figure 4.1. We have \( f(0, 0) = 0 \), while there exist negative values of \( f \) in any \( \varepsilon \)-neighborhood about the point \( (0, 0) \). In fact, taking \( d = (\sin \theta, \cos \theta)' \), we have
\[
 f(\bar{x} + \lambda d) - f(\bar{x}) = 2\sin^2 \theta \lambda^2 - 3\sin \theta \cos^2 \theta \lambda^3 + \cos^4 \theta \lambda^4; 
\]
and for this to be nonnegative for all \( -\delta_\theta \leq \lambda \leq \delta_\theta \), \( \delta_\theta > 0 \), we observe that as \( \theta \to 0^+ \), we get \( \delta_\theta \to 0^+ \) as well (see Exercise 4.11), although at \( \theta = 0 \) we get \( \delta_\theta = \infty \). Hence, we cannot derive a \( \delta > 0 \) such that \( f(\bar{x} + \lambda d) - f(\bar{x}) \geq 0 \), for all \( d \in \mathbb{R}^n \) and \( -\delta \leq \lambda \leq \delta \), so \( \bar{x} \) is not a local minimum.
Figure 4.1 Regions of zero, positive, and negative values of 
\[ f(x_1, x_2) = (x_2^2 - x_1)(x_2^2 - 2x_1). \]

To afford further insight into the multivariate case, let us examine a situation in which \( f: \mathbb{R}^n \to \mathbb{R} \) is twice continuously differentiable, and at a given point \( \bar{x} \in \mathbb{R}^n \), we have that \( \nabla f(\bar{x}) = 0 \) but \( \mathbf{H}(\bar{x}) \) is indefinite. Hence, there exist directions \( d_1 \) and \( d_2 \) in \( \mathbb{R}^n \) such that \( d_1^T \mathbf{H}(\bar{x}) d_1 > 0 \) and \( d_2^T \mathbf{H}(\bar{x}) d_2 < 0 \). Defining 
\[ F(\bar{x}, d_j)(\lambda) = f(\bar{x} + \lambda d_j) \] for \( j = 1, 2 \), and denoting derivatives by primes, we get 
\[ F_{d_1}^j(\lambda) = \nabla f(\bar{x} + \lambda d_j)^T d_j \] and \[ F_{d_2}^j(\lambda) = d_j^T \mathbf{H}(\bar{x} + \lambda d_j) d_j \] for \( j = 1, 2 \).

Hence, for \( j = 1 \), we have \( F_{d_1}^j(0) = 0 \), \( F_{d_2}^j(0) > 0 \); and moreover, by continuity of the second derivative, \( F_{d_1}^j(\lambda) > 0 \), for \( |\lambda| \) sufficiently small. Hence, \( F_{d_1}(\lambda) \) is strictly convex in some \( \epsilon \)-neighborhood of \( \lambda = 0 \), achieving a strict local minimum at \( \lambda = 0 \). Similarly, for \( j = 2 \), noting that \( F_{d_2}^j(0) = 0 \) and \( F_{d_2}^j(0) < 0 \), we conclude that \( F_{d_2}(\lambda) \) is strictly concave in some \( \epsilon \)-neighborhood of \( \lambda = 0 \), achieving a strict local maximum at \( \lambda = 0 \). Hence, as foretold by Theorem 4.1.3, \( \bar{x} = 0 \) is neither a local minimum nor a local maximum. Such a point \( \bar{x} \) is called a saddle point (or an inflection point). Figure 4.2 illustrates the situation. Observe the convex and concave cross sections of the function in the respective directions \( d_1 \) and \( d_2 \) about the point \( \bar{x} \) at which \( \nabla f(\bar{x}) = 0 \), which gives the function the appearance of a saddle in the vicinity of \( \bar{x} \).
4.1.7 Examples

Example 1: Univariate Function  
To illustrate the necessary and sufficient conditions of this section, consider the problem to minimize \( f(x) = (x^2 - 1)^3 \).

First, let us determine the candidate points for optimality satisfying the first-order necessary condition that \( \nabla f(x) = 0 \). Note that \( \nabla f(x) = f'(x) = 6x(x^2 - 1)^2 = 0 \) when \( x = 0, 1, \) or \(-1\). Hence, our candidate points for local optimality are \( x = 0, 1, \) or \(-1\). Now let us examine the second-order derivatives. We have \( H(x) = f''(x) = 24x(x^2 - 1) + 6(x^2 - 1)^2 \), and hence \( H(1) = H(-1) = 0 \) and \( H(0) = 6 \). Since \( H \) is positive definite at \( x = 0 \), we have by Theorem 4.1.4 that \( x = 0 \) is a strict local minimum. However, at \( x = +1 \) or \(-1\), \( H \) is both positive and negative semidefinite; and although it satisfies the second-order necessary condition of Theorem 4.1.3, this is not sufficient for us to conclude anything about the behavior of \( f \) at these points. Hence, we continue and examine the third-order derivative \( f'''(x) = 48x(x^2 - 1) + 48x^3 + 24x(x^2 - 1) \). Evaluating this at the two candidate points \( x = \pm 1 \) in question, we obtain \( f''(1) > 0 \) and \( f''(-1) < 0 \). By Theorem 4.1.6 it follows that we have neither a local minimum nor a local maximum at these points, and these points are merely inflection points.

Example 2: Multivariate Function  
Consider the bivariate function \( f(x_1, x_2) = x_1^3 + x_2^3 \). Evaluating the gradient and the Hessian of \( f \), we obtain

\[
\nabla f(x) = \begin{pmatrix} 3x_1^2 \\ 3x_2^2 \end{pmatrix} \quad \text{and} \quad H(x) = \begin{bmatrix} 6x_1 & 0 \\ 0 & 6x_2 \end{bmatrix}.
\]

The first-order necessary condition \( \nabla f(x) = 0 \) yields \( x = (0, 0)^T \) as the single candidate point. However, \( H(x) \) is the zero matrix; and although it satisfies the second-order necessary condition of Theorem 4.1.3, we need to examine higher-order derivatives to make any conclusive statement about the point \( x \). Defining \( F_{(x,d)}(\lambda) = f(x + \lambda d) = F_d(\lambda) \), say, we have \( F_d'(\lambda) = \nabla f(x + \lambda d)'d \), \( F_d''(\lambda) = \nabla f(x + \lambda d)''d \), \( F_d'''(\lambda) = \nabla f(x + \lambda d)'''d \), \( F_d(iv)'d \), and so on. Figure 4.2 Saddle point at \( x \).
\[ d' H(\bar{x} + \lambda d) d, \text{ and } F_d^*(\lambda) = \sum_{i=1}^{n} \sum_{j=1}^{m} d_i d_j d_k f_{ijk}(\bar{x} + \lambda d). \] Noting that \( f_{111}(x) = 6, f_{222}(x) = 6, \text{ and } f_{ijk}(x) = 0 \) otherwise, we obtain \( F_d^*(0) = 6d_1^3 + 6d_2^3. \) Since there exist directions \( d \) for which the first nonzero derivative term at \( \lambda = 0 \) is \( F_d^*(0), \) which is of odd order, \( \bar{x} = (0,0)^t \) is an inflection point and is therefore neither a local minimum nor a local maximum. In fact, note that \( F_d^*(\lambda) = 6\lambda(d_1^3 + d_2^3) \) can be made to take on opposite signs about \( \lambda = 0 \) along any direction \( d \) for which \( d_1^3 + d_2^3 \neq 0; \) so the function switches from a convex to a concave function, or vice versa, about the point \( 0 \) along any direction \( d. \) Observe also that \( H \) is positive semidefinite over \( \{x : x_1 \geq 0, x_2 \geq 0\}; \) and hence, over this region, the function is convex, yielding \( \bar{x} = (0,0)^t \) as a global minimum. Similarly, \( \bar{x} = (0,0)^t \) is a global maximum over the region \( \{x : x_1 \leq 0, x_2 \leq 0\}. \)

4.2 Problems Having Inequality Constraints

In this section we first develop a necessary optimality condition for the problem to minimize \( f(x) \) subject to \( x \in S \) for a general set \( S. \) Later, we let \( S \) be more specifically defined as the feasible region of a nonlinear programming problem of the form to minimize \( f(x) \) subject to \( g(x) \leq 0 \) and \( x \in X. \)

Geometric Optimality Conditions

In Theorem 4.2.2 we develop a necessary optimality condition for the problem to minimize \( f(x) \) subject to \( x \in S, \) using the cone of feasible directions defined below.

4.2.1 Definition

Let \( S \) be a nonempty set in \( \mathbb{R}^n, \) and let \( \bar{x} \in \text{cl} \ S. \) The cone of feasible directions of \( S \) at \( \bar{x}, \) denoted by \( D, \) is given by

\[ D = \{d : d \neq 0, \text{ and } \bar{x} + \lambda d \in S \text{ for all } \lambda \in (0,\delta) \text{ for some } \delta > 0\}. \]

Each nonzero vector \( d \in D \) is called a feasible direction. Moreover, given a function \( f : \mathbb{R}^n \to \mathbb{R}, \) the cone of improving directions at \( \bar{x}, \) denoted by \( F, \) is given by

\[ F = \{d : f(\bar{x} + \lambda d) < f(\bar{x}) \text{ for all } \lambda \in (0,\delta) \text{ for some } \delta > 0\}. \]

Each direction \( d \in F \) is called an improving direction, or a descent direction, of \( f \) at \( \bar{x}. \)

From the above definitions, it is clear that a small movement from \( \bar{x} \) along a vector \( d \in D \) leads to feasible points, whereas a similar movement along a \( d \in F \) vector leads to solutions of improving objective value. Furthermore,
from Theorem 4.1.2, if $\nabla f(\bar{x})^t d < 0$, $d$ is an improving direction; that is, starting from $\bar{x}$, a small movement along $d$ will reduce the value of $f$. As shown in Theorem 4.2.2, if $\bar{x}$ is a local minimum and if $\nabla f(\bar{x})^t d < 0$, $d \notin D$, that is, a necessary condition for local optimality is that every improving direction is not a feasible direction. This fact is illustrated in Figure 4.3, where the vertices of the cones $F_0 \equiv \{d : \nabla f(\bar{x})^t d < 0\}$ and $D$ are translated from the origin to $\bar{x}$ for convenience.

4.2.2 Theorem

Consider the problem to minimize $f(x)$ subject to $x \in S$, where $f : \mathbb{R}^n \to \mathbb{R}$ and $S$ is a nonempty set in $\mathbb{R}^n$. Suppose that $f$ is differentiable at a point $\bar{x} \in S$. If $\bar{x}$ is a local optimal solution, $F_0 \cap D = \emptyset$, where $F_0 = \{d : \nabla f(\bar{x})^t d < 0\}$ and $D$ is the cone of feasible directions of $S$ at $\bar{x}$. Conversely, suppose that $F_0 \cap D = \emptyset$, $f$ is pseudoconvex at $\bar{x}$, and that there exists an $\varepsilon$-neighborhood $N_\varepsilon(\bar{x})$, $\varepsilon > 0$, such that $d = (x - \bar{x}) \in D$ for any $x \in S \cap N_\varepsilon(\bar{x})$. Then $\bar{x}$ is a local minimum of $f$.

Proof

By contradiction, suppose that there exists a vector $d \in F_0 \cap D$. Then, by Theorem 4.1.2, there exists a $\delta_1 > 0$ such that

$$f(\bar{x} + \lambda d) < f(\bar{x}) \quad \text{for each } \lambda \in (0, \delta_1). \quad (4.5a)$$

Figure 4.3 Necessary condition $F_0 \cap D = \emptyset$. 
Furthermore, by Definition 4.2.1, there exists a $\delta_2 > 0$ such that

$$\overline{x} + \lambda d \in S \quad \text{for each } \lambda \in (0, \delta_2). \quad (4.5b)$$

The assumption that $\overline{x}$ is a local optimal solution to the problem is not compatible with (4.5). Thus, $F_0 \cap D = \emptyset$.

Conversely, suppose that $F_0 \cap D = \emptyset$ and that the given conditions in the converse statement of the theorem hold true. Then we must have $f(x) \geq f(\overline{x})$ for all $x \in S \cap N_\varepsilon(\overline{x})$. To understand this, suppose that $f(\hat{x}) < f(\overline{x})$ for some $\hat{x} \in S \cap N_\varepsilon(\overline{x})$. By the assumption on $S \cap N_\varepsilon(\overline{x})$, we have $d = (\hat{x} - \overline{x}) \in D$. Moreover, by the pseudoconvexity of $f$ at $\overline{x}$, we have that $\nabla f(\overline{x})'d < 0$; or else, if $\nabla f(\overline{x})'d \geq 0$, we would obtain $f(\hat{x}) = f(\overline{x} + d) \geq f(\overline{x})$. We have therefore shown that if $\overline{x}$ is not a local minimum over $S \cap N_\varepsilon(\overline{x})$, there exists a direction $d \in F_0 \cap D$, which is a contradiction. This completes the proof.

Observe that the set $F_0$ defined in Theorem 4.2.2 provides an algebraic characterization for the set of improving directions $F$. In fact, we have $F_0 \subseteq F$ in general by Theorem 4.1.2. Also, if $d \in F$, we must have $\nabla f(\overline{x})'d \leq 0$, or else, analogous to Theorem 4.1.2, $\nabla f(\overline{x})'d > 0$ would imply that $d$ is an ascent direction. Hence, we have

$$F_0 \subseteq F \subseteq F_0' = \{d \neq 0 : \nabla f(\overline{x})'d \leq 0\}. \quad (4.6)$$

Note that when $\nabla f(\overline{x})'d = 0$, we are unsure about the behavior of $f$ as we proceed from $\overline{x}$ along the direction $d$, unless we know more about the function. For example, it might very well be that $\nabla f(\overline{x}) = 0$, and there might exist directions of motion that give descent or ascent, or even hold the value of $f$ constant as we move away from $\overline{x}$. Hence, it is entirely possible to have $F_0 \subset F \subset F_0'$ (see Figure 4.1, for example). However, if $f$ is pseudoconvex, we know that whenever $\nabla f(\overline{x})'d \geq 0$, we have $f(\overline{x} + \lambda d) \geq f(\overline{x})$ for all $\lambda \geq 0$. Hence, if $f$ is pseudoconvex, $d \in F$ implies that $d \in F_0$ as well, so from (4.6), we have $F_0 = F$. Similarly, if $f$ is strictly pseudoconcave, we know that whenever $d \in F_0'$, we have $f(\overline{x} + \lambda d) < f(\overline{x})$ for all $\lambda > 0$, so we have $d \in F$ as well. Consequently, we obtain $F = F_0'$ in this case. This establishes the following result, stated in terms of the weaker assumption of pseudoconvexity or strict pseudoconcavity at $\overline{x}$ itself, rather than everywhere.
4.2.3 Lemma

Consider a differentiable function $f: \mathbb{R}^n \to \mathbb{R}$, and let $F, F_0, F_0'$ be as defined in Definition 4.2.1, Theorem 4.2.2, and (4.6), respectively. Then we have $F_0 \subseteq F \subseteq F_0'$. Moreover, if $f$ is pseudoconvex at $\bar{x}$, $F = F_0$, and if $f$ is strictly pseudoconcave at $\bar{x}$, $F = F_0'$.

We now specify the feasible region $S$ as follows:

$$S = \{ x \in X : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m \},$$

where $g_i: \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$ and $X$ is a nonempty open set in $\mathbb{R}^n$. This gives us the following inequality constrained nonlinear programming problem:

$$P: \text{Minimize } f(x)$$
subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$

$x \in X$.

Recall that a necessary condition for local optimality at $\bar{x}$ is that $F_0 \cap D = \emptyset$, where $F_0$ is an open half-space defined in terms of the gradient vector $\nabla f(\bar{x})$, and $D$ is the cone of feasible directions, which is not necessarily defined in terms of the gradients of the functions involved. This precludes us from converting the geometric optimality condition $F_0 \cap D = \emptyset$ into a more usable algebraic statement involving equations or inequalities. As Lemma 4.2.4 indicates, we can define an open cone $G_0$ in terms of the gradients of the binding constraints at $\bar{x}$ such that $G_0 \subseteq D$. Since $F_0 \cap D = \emptyset$ must hold at $\bar{x}$ and since $G_0 \subseteq D$, $F_0 \cap G_0 = \emptyset$ is also a necessary optimality condition. Since $F_0$ and $G_0$ are both defined in terms of the gradient vectors, we use the condition $F_0 \cap G_0 = \emptyset$ later in the section to develop the optimality conditions credited to Fritz John. With mild additional assumptions, the conditions reduce to the well-known Karush–Kuhn–Tucker (KKT) optimality conditions.

4.2.4 Lemma

Consider the feasible region $S = \{ x \in X : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m \}$, where $X$ is a nonempty open set in $\mathbb{R}^n$ and where $g_i: \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$. Given a feasible point $\bar{x} \in S$, let $I = \{ i : g_i(\bar{x}) = 0 \}$ be the index set for the binding, or active, or tight constraints, and assume that $g_i$ for $i \in I$ are differentiable at $\bar{x}$ and that $g_i$ for $i \notin I$ are continuous at $\bar{x}$. Define the sets

$$G_0 = \{ d : \nabla g_i(\bar{x})^t d < 0 \text{ for each } i \in I \}$$

$$G_0' = \{ d \neq 0 : \nabla g_i(\bar{x})^t d \leq 0 \text{ for each } i \in I \}.$$
Then we have

\[ G_0 \subseteq D \subseteq G'_0. \] (4.7)

Moreover, if \( g_i, i \in I \), are strictly pseudoconvex at \( \bar{x} \), \( D = G_0 \); and if \( g_i, i \in I \), are pseudoconcave at \( \bar{x} \), \( D = G'_0 \).

**Proof**

Let \( d \in G_0 \). Since \( \bar{x} \in X \), and \( X \) is open, there exists a \( \delta_1 > 0 \) such that

\[ \bar{x} + \lambda d \in X \quad \text{for} \quad \lambda \in (0, \delta_1). \] (4.8a)

Also, since \( g_i(\bar{x}) < 0 \) and \( g_i \) is continuous at \( \bar{x} \) for \( i \not\in I \), there exists a \( \delta_2 > 0 \) such that

\[ g_i(\bar{x} + \lambda d) < 0 \quad \text{for} \quad \lambda \in (0, \delta_2) \quad \text{and} \quad i \not\in I. \] (4.8b)

Furthermore, since \( d \in G_0 \), \( \nabla g_i(\bar{x})'d < 0 \) for each \( i \in I \), and by Theorem 4.1.2, there exists a \( \delta_3 > 0 \) such that

\[ g_i(\bar{x} + \lambda d) < g_i(\bar{x}) = 0 \quad \text{for} \quad \lambda \in (0, \delta_3) \quad \text{and} \quad i \in I. \] (4.8c)

From (4.8a, b, c), it is clear that points of the form \( \bar{x} + \lambda d \) are feasible to \( S \) for each \( \lambda \in (0, \delta) \), where \( \delta = \min\{\delta_1, \delta_2, \delta_3\} > 0 \). Thus, \( d \in D \), where \( D \) is the cone of feasible directions of the feasible region at \( \bar{x} \). We have shown thus far that \( d \in G_0 \) implies that \( d \in D \), and hence \( G_0 \subseteq D \).

Similarly, if \( d \in D \), we must have \( d \in G'_0 \), since otherwise, if \( \nabla g_i(\bar{x})'d > 0 \) for any \( i \in I \), we would obtain via Theorem 4.1.2 that \( g_i(\bar{x} + \lambda d) > g_i(\bar{x}) = 0 \) for all \( |\lambda| \) sufficiently small, contradicting the hypotheses that \( d \in D \). Hence, \( D \subseteq G'_0 \). This establishes (4.7).

Now suppose that \( g_i, i \in I \), are strictly pseudoconvex at \( \bar{x} \), and let \( d \in D \).

Then we must have \( d \in G_0 \) as well, because otherwise, if \( \nabla g_i(\bar{x})'d \geq 0 \) for any \( i \in I \), we would obtain that \( g_i(\bar{x} + \lambda d) > g_i(\bar{x}) = 0 \) for all \( \lambda > 0 \), contradicting that \( d \in D \). Hence, from (4.7), we get \( D = G_0 \) in this case.

Finally, suppose that \( g_i, i \in I \), are pseudoconcave at \( \bar{x} \), and consider any \( d \in G'_0 \). We therefore have \( g_i(\bar{x} + \lambda d) \leq g_i(\bar{x}) = 0 \) for all \( \lambda \geq 0 \) for each \( i \in I \). Moreover, by the continuity of \( g_i, i \not\in I \), and since \( X \) is an open set, we obtain as above that \( (\bar{x} + \lambda d) \in S \) for all \( |\lambda| \) sufficiently small, so \( d \in D \). This establishes that \( G'_0 \subseteq D \), so from (4.7) we obtain \( D = G'_0 \) in this case. This completes the proof.
As an illustration, note that in Figure 4.3, $G_0 = D \subset G_0'$, whereas in Figure 2.18 with $\bar{x} = (0,0)^t$, since the constraint functions are affine, we have $G_0' \subset D = G_0'$.

Lemma 4.2.4 leads directly to the following result.

### 4.2.5 Theorem

Consider Problem P to minimize $f(x)$ subject to $x \in X$ and $g_i(x) \leq 0$ for $i = 1, \ldots, m$, where $X$ is a nonempty open set in $\mathbb{R}^n$, $f: \mathbb{R}^n \to \mathbb{R}$, and $g_i: \mathbb{R}^n \to \mathbb{R}$, for $i = 1, \ldots, m$. Let $\bar{x}$ be a feasible point, and denote $I = \{i: g_i(\bar{x}) = 0\}$. Furthermore, suppose that $f$ and $g_i$ for $i \in I$ are differentiable at $\bar{x}$ and that $g_i$ for $i \notin I$ are continuous at $\bar{x}$. If $\bar{x}$ is a local optimal solution, $F_0 \cap G_0 = \emptyset$, and $G_0 = \{d: Vg_i(\bar{x})'d < 0 \text{ for each } i \in I\}$. Conversely, if $F_0 \cap G_0 = \emptyset$, and $f$ and $g_i$, $i \in I$, are as specified in the theorem. Then, redefining the feasible region $S$ only in terms of the binding constraints by dropping the nonbinding constraints, we have that $G_0 = D$ by Lemma 4.2.4, so we conclude that $F_0 \cap D = \emptyset$. Furthermore, since the level sets $g_i(x) \leq 0$, for $i \in I$, are convex over some $\epsilon$-neighborhood $N_\epsilon(\bar{x})$ of $\bar{x}$, $\epsilon > 0$, it follows that $S \cap N_\epsilon(\bar{x})$ is a convex set. Since we also have $F_0 \cap D = \emptyset$ from above, and since $f$ is pseudoconvex at $\bar{x}$ and $g_i$, $i \in I$, are strictly pseudoconvex over some $\epsilon$-neighborhood of $\bar{x}$, $\bar{x}$ is a local minimum.

**Proof**

Let $\bar{x}$ be a local minimum. Then we have the following string of implications via Theorem 4.2.2 and (4.7) of Lemma 4.2.4, which proves the first part of the theorem:

\[
\bar{x} \text{ is a local minimum} \Rightarrow F_0 \cap D = \emptyset \Rightarrow F_0 \cap G_0 = \emptyset. \tag{4.9a}
\]

Conversely, suppose that $F_0 \cap G_0 = \emptyset$ and that $f$ and $g_i$, $i \in I$, are as specified in the theorem. Then, redefining the feasible region $S$ only in terms of the binding constraints by dropping the nonbinding constraints, we have that $G_0 = D$ by Lemma 4.2.4, so we conclude that $F_0 \cap D = \emptyset$. Furthermore, since the level sets $g_i(x) \leq 0$, for $i \in I$, are convex over some $\epsilon$-neighborhood $N_\epsilon(\bar{x})$ of $\bar{x}$, $\epsilon > 0$, it follows that $S \cap N_\epsilon(\bar{x})$ is a convex set. Since we also have $F_0 \cap D = \emptyset$ from above, and since $f$ is pseudoconvex at $\bar{x}$, we conclude from the converse statement of Theorem 4.2.2 that $\bar{x}$ is a local minimum. This statement continues to hold true by including the nonbinding constraints within $S$, and this completes the proof.

Observe that under the converse hypothesis of Theorem 4.2.5, and assuming that $g_i$, $i \notin I$, are continuous at $\bar{x}$, we have, noting (4.9a),

\[
\bar{x} \text{ is a local minimum} \iff F_0 \cap D = \emptyset \iff F_0 \cap G_0 = \emptyset. \tag{4.9b}
\]

There is a useful insight worth deriving at this point. Note from Definition 4.2.1 that if $\bar{x}$ is a local minimum, then clearly we must have $F \cap D = \emptyset$. However, the converse is not necessarily true. That is, if $F \cap D = \emptyset$, this does not necessarily imply that $\bar{x}$ is a local minimum. For example, if
\[ S = \{ x = (x_1, x_2) : x_2 = x_1^2 \} \] and if \( f(x) = x_2 \), the point \( \bar{x} = (1,1)' \) is clearly not a local minimum, since \( f \) can be decreased by reducing \( x_1 \). However, for the given point \( \bar{x} \), the set \( D = \emptyset \), since no straight-line directions can lead to feasible solutions, whereas improving feasible solutions are accessible via curvilinear directions. Hence, \( F \cap D = \emptyset \), but \( \bar{x} \) is not a local minimum. But now, if \( f \) is pseudoconvex at \( \bar{x} \), and if there exists an \( \varepsilon > 0 \) such that for any \( x \in S \cap N_\varepsilon(\bar{x}) \), we have \( d = (x - \bar{x}) \in D \) [as, e.g., if \( S \cap N_\varepsilon(\bar{x}) \) is a convex set], \( F_0 = F \) by Lemma 4.2.3; and noting (4.9a) and the converse to Theorem 4.2.2, we obtain in this case

\[
F \cap D = \emptyset \iff F_0 \cap D = \emptyset \iff \bar{x} \text{ is a local minimum.}
\]

4.2.6 Example

Minimize \((x_1 - 3)^2 + (x_2 - 2)^2\)
subject to \(x_1^2 + x_2^2 \leq 5\)
\(x_1 + x_2 \leq 3\)
\(x_1 \geq 0\)
\(x_2 \geq 0\).

In this case we let \(g_1(x) = x_1^2 + x_2^2 - 5\), \(g_2(x) = x_1 + x_2 - 3\), \(g_3(x) = -x_1\), \(g_4(x) = -x_2\), and \(X = \mathbb{R}^2\). Consider the point \(\bar{x} = (9/5, 6/5)'\), and note that the only binding constraint is \(g_2(x) = x_1 + x_2 - 3\). Also, note that

\[
\nabla f(\bar{x}) = \left( \frac{-12}{5}, \frac{-8}{5} \right)'
\]
and
\[
\nabla g_2(\bar{x}) = (1,1)'.
\]

Figure 4.4 \(F_0 \cap G_0 \neq \emptyset\) at a nonoptimal point.
The Fritz John and Karush-Kuhn-Tucker Optimality Conditions

The sets $F_0$ and $G_0$, with the origin translated to $(9/5,6/5)'$ for convenience, are shown in Figure 4.4. Since $F_0 \cap G_0 \neq \emptyset$, $\bar{x} = (9/5,6/5)'$ is not a local optimal solution to the above problem.

Now consider the point $\bar{x} = (2,1)'$, and note that the first two constraints are binding. The corresponding gradients at this point are

$$\nabla f(\bar{x}) = (-2,-2)', \quad \nabla g_1(\bar{x}) = (4,2)', \quad \nabla g_2(\bar{x}) = (1,1)'.$$

The sets $F_0$ and $G_0$ are shown in Figure 4.5, and indeed, $F_0 \cap G_0 = \emptyset$. Note also that the sufficiency condition of Theorem 4.2.5 is not satisfied because $g_2$ is not strictly pseudoconvex over any neighborhood about $\bar{x}$. However, from Figure 4.5 we observe that we also have $F_0 \cap G_0' = \emptyset$ in this case; so by (4.7) we have $F_0 \cap D = \emptyset$. By the converse to Theorem 4.2.2, we can now conclude that $\bar{x}$ is a local minimum. In fact, since the problem is a convex program with a strictly convex objective function, this in turn implies that $\bar{x}$ is the unique global minimum.

It might be interesting to note that the utility of Theorem 4.2.5 also depends on how the constraint set is expressed. This is illustrated by Example 4.2.7.

**4.2.7 Example**

Minimize $$(x_1 - 1)^2 + (x_2 - 1)^2$$ subject to $$(x_1 + x_2 - 1)^3 \leq 0$$ $$x_1 \geq 0$$ $$x_2 \geq 0.$$ 

Note that the necessary condition of Theorem 4.2.5 holds true at each feasible point with $x_1 + x_2 = 1$. However, the constraint set can be represented equivalently by

![Figure 4.5](image) $F_0 \cap G_0 = \emptyset$ at an optimal point.
It can easily be verified that $F_0 \cap G_0 = \emptyset$ is now satisfied only at the point $(1/2, 1/2)$. Moreover, it can also easily be verified that $F_0 \cap G'_0 = \emptyset$ in this case, so by (4.7), $F_0 \cap D = \emptyset$. Following the converse to Theorem 4.2.2 and noting the convexity of the feasible region and the strict convexity of the objective function, we can conclude that $\bar{x} = (1/2, 1/2)'$ is indeed the unique global minimum to the problem.

There are several cases where the necessary conditions of Theorem 4.2.5 are satisfied trivially by possibly nonoptimal points also. Some of these cases are discussed below.

Suppose that $\bar{x}$ is a feasible point such that $\nabla f(\bar{x}) = 0$. Clearly, $F_0 = \{d : \nabla f(\bar{x})'d < 0\} = \emptyset$ and hence $F_0 \cap G_0 = \emptyset$. Thus, any point $\bar{x}$ having $\nabla f(\bar{x}) = 0$ satisfies the necessary optimality conditions. Similarly, any point $\bar{x}$ having $\nabla g_i(\bar{x}) = 0$ for some $i \in I$ will also satisfy the necessary conditions. Now consider the following example with an equality constraint:

\[
\text{Minimize } f(x) \\
\text{subject to } g(x) = 0.
\]

The equality constraint $g(x) = 0$ could be replaced by the inequality constraints $g_1(x) = g(x) \leq 0$ and $g_2(x) = -g(x) \leq 0$. Let $\bar{x}$ be any feasible point. Then $g_1(\bar{x}) = g_2(\bar{x}) = 0$. Note that $\nabla g_1(\bar{x}) = -\nabla g_2(\bar{x})$, and therefore there could exist no vector $d$ such that $\nabla g_1(\bar{x})'d < 0$ and $\nabla g_2(\bar{x})'d < 0$. Therefore, $G_0 = \emptyset$ and hence $F_0 \cap G_0 = \emptyset$. In other words, the necessary condition of Theorem 4.2.5 is satisfied by all feasible solutions and is hence not usable.

**Fritz John Optimality Conditions**

We now reduce the geometric necessary optimality condition $F_0 \cap G_0 = \emptyset$ to a statement in terms of the gradients of the objective function and of the binding constraints. The resulting optimality conditions, credited to Fritz John [1948], are given below.

**4.2.8 Theorem (Fritz John Necessary Conditions)**

Let $X$ be a nonempty open set in $\mathbb{R}^n$ and let $f : \mathbb{R}^n \to \mathbb{R}$, and $g_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, ..., m$. Consider Problem P to minimize $f(x)$ subject to $x \in X$ and $g_i(x) \leq 0$ for $i = 1, ..., m$. Let $\bar{x}$ be a feasible solution, and denote $I = \{i : g_i(\bar{x}) = 0\}$. Furthermore, suppose that $f$ and $g_i$ for $i \in I$ are differentiable at $\bar{x}$ and that $g_i$ for $i \not\in I$
are continuous at \( \bar{x} \). If \( \bar{x} \) solves Problem P locally, there exist scalars \( u_0 \) and \( u_i \) for \( i \in I \) such that

\[
0 = u_0 \nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) \quad \text{with} \quad u_0, u_i \geq 0 \quad \text{for} \ i \in I \quad (u_0, u_I) \neq (0, 0),
\]

where\( \mathbf{u}_I \) is the vector whose components are \( u_i \) for \( i \in I \). Furthermore, if \( g_i \) for \( i \notin I \) are also differentiable at \( \bar{x} \), the foregoing conditions can be written in the following equivalent form:

\[
0 = u_0 \nabla f(\bar{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) \quad \text{with} \quad u_i g_i(\bar{x}) = 0 \quad \text{for} \ i = 1, \ldots, m \quad \text{and} \quad (u_0, u) \neq (0, 0),
\]

where \( u \) is the vector whose components are \( u_i \) for \( i = 1, \ldots, m \).

**Proof**

Since \( \bar{x} \) solves Problem P locally, by Theorem 4.2.5 there exists no vector \( \mathbf{d} \) such that \( \nabla f(\bar{x})' \mathbf{d} < 0 \) and \( \nabla g_i(\bar{x})' \mathbf{d} < 0 \) for each \( i \in I \). Now let \( \mathbf{A} \) be the matrix whose rows are \( \nabla f(\bar{x})' \) and \( \nabla g_i(\bar{x})' \) for \( i \in I \). The necessary optimality condition of Theorem 4.2.5 is then equivalent to the statement that the system \( \mathbf{A} \mathbf{d} < 0 \) is inconsistent. By Gordan's theorem 2.4.9 there exists a nonzero vector \( \mathbf{p} \geq 0 \) such that \( \mathbf{A}' \mathbf{p} = 0 \). Denoting the components of \( \mathbf{p} \) by \( u_0 \) and \( u_i \) for \( i \in I \), the first part of the result follows. The equivalent form of the necessary conditions is readily obtained by letting \( u_i = 0 \) for \( i \notin I \), and the proof is complete.

Pertaining to the conditions of Theorem 4.2.8, the scalars \( u_0 \) and \( u_i \) for \( i = 1, \ldots, m \) are called Lagrangian, or Lagrange, multipliers. The condition that \( \bar{x} \) be feasible to Problem P is called the primal feasibility (PF) condition, whereas the requirements \( u_0 \nabla f(\bar{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) = 0, \ (u_0, \mathbf{u}) \geq (0, 0), \) and \( (u_0, \mathbf{u}) \neq (0, 0) \) are sometimes referred to as dual feasibility (DF) conditions. The condition \( u_i g_i(\bar{x}) = 0 \) for \( i = 1, \ldots, m \) is called the complementary slackness (CS) condition. It requires that \( u_i = 0 \) if the corresponding inequality is nonbinding, that is, if \( g_i(\bar{x}) < 0 \). Similarly, it permits \( u_i > 0 \) only for those constraints that are binding.
Together, the PF, DF, and the CS conditions are called the Fritz John (FJ) optimality conditions. Any point $\mathbf{x}$ for which there exist Lagrangian multipliers $(u_0, \mathbf{u})$ such that $(\mathbf{x}, u_0, \mathbf{u})$ satisfy the FJ conditions is called a Fritz John point. The Fritz John conditions can also be written in vector notation as follows, in addition to the PF requirement:

$$u_0 \nabla f(\mathbf{x}) + \nabla g(\mathbf{x})' \mathbf{u} = 0$$
$$\mathbf{u}' g(\mathbf{x}) = 0$$
$$(u_0, \mathbf{u}) \geq (0, 0)$$
$$(u_0, \mathbf{u}) \neq (0, 0).$$

Here, $\nabla g(\mathbf{x})$ is an $m \times n$ Jacobian matrix whose $i$th row is $\nabla g_i(\mathbf{x})'$, and $\mathbf{u}$ is an $m$-vector denoting the Lagrangian multipliers.

### 4.2.9 Example

Minimize $(x_1 - 3)^2 + (x_2 - 2)^2$
subject to $x_1^2 + x_2^2 \leq 5$
$x_1 + 2x_2 \leq 4$
$-x_1 \leq 0$
$-x_2 \leq 0.$

The feasible region for the above problem is illustrated in Figure 4.6. We now verify that the Fritz John conditions hold true at the optimal point $(2, 1)$. First, note that the set of binding constraints $I$ at $\mathbf{x} = (2, 1)'$ is given by $I = \{1, 2\}$. Thus, the Lagrangian multipliers $u_3$ and $u_4$ associated with $-x_1 \leq 0$ and $-x_2 \leq 0$, respectively, are equal to zero. Note that

![Figure 4.6 Feasible Region in Example 4.2.9.](image-url)
The Fritz John and Karush–Kuhn–Tucker Optimality Conditions

\[ \nabla f(\bar{x}) = (-2, -2)' , \quad \nabla g_1(\bar{x}) = (4, 2)' , \quad \nabla g_2(\bar{x}) = (1, 2)' . \]

Hence, to satisfy the Fritz John conditions, we now need a nonzero vector \((u_0, u_1, u_2) \geq 0\) satisfying

\[
\begin{bmatrix}
-u_0 \\
-u_1 \\
u_2
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}.
\]

This implies that \(u_1 = u_0/3\) and \(u_2 = 2u_0/3\). Taking \(u_1\) and \(u_2\) as such for any \(u_0 > 0\), we satisfy the FJ conditions. As another illustration, let us check whether the point \(\hat{x} = (0, 0)'\) is a FJ point. Here the set of binding constraints is \(I = \{3, 4\}\), and thus \(u_1 = u_2 = 0\). Note that

\[ \nabla f(\hat{x}) = (-6, -4)' , \quad \nabla g_3(\hat{x}) = (-1, 0)' , \quad \nabla g_4(\hat{x}) = (0, -1)' . \]

Also, note that the DF condition

\[
\begin{bmatrix}
-u_0 \\
-u_3 \\
u_4
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}
\]

holds true if and only if \(u_3 = -6u_0\) and \(u_4 = -4u_0\). If \(u_0 > 0\), \(u_3\) and \(u_4\) are negative, contradicting the nonnegativity restrictions. If, on the other hand, \(u_0 = 0\), then \(u_3 = u_4 = 0\), which contradicts the stipulation that the vector \((u_0, u_3, u_4)\) is nonzero. Thus, the Fritz John conditions do not hold true at \(\hat{x} = (0, 0)'\), which also shows that the origin is not a local optimal point.

4.2.10 Example

Consider the following problem from Kuhn and Tucker [1951]:

Minimize \(-x_1\)

subject to \(x_2 - (1 - x_1)^3 \leq 0\)

\[-x_2 \leq 0.\]

The feasible region is illustrated in Figure 4.7. We now verify that the Fritz John conditions indeed hold true at the optimal point \(\bar{x} = (1, 0)'\). Note that the set of binding constraints at \(\bar{x}\) is given by \(I = \{1, 2\}\). Also,

\[ \nabla f(\bar{x}) = (-1, 0)' , \quad \nabla g_1(\bar{x}) = (0, 1)' , \quad \nabla g_2(\bar{x}) = (0, -1)' . \]

The DF condition

\[
\begin{bmatrix}
-u_0 \\
u_1 \\
u_2
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}
\]
is true only if $u_0 = 0$. Thus, the Fritz John conditions hold true at $\bar{x}$ by letting $u_0 = 0$ and $u_1 = u_2 = \alpha$, where $\alpha$ is any positive scalar.

**4.2.11 Example**

Minimize $-x_1$

subject to $x_1 + x_2 - 1 \leq 0$

$-x_2 \leq 0$.

The feasible region is sketched in Figure 4.8, and the optimal point is $\bar{x} = (1,0)^f$. Note that

$\nabla f(\bar{x}) = (-1,0)^f$, \quad $\nabla g_1(\bar{x}) = (1,1)^f$, \quad $\nabla g_2(\bar{x}) = (0,-1)^f$,

and the Fritz John conditions hold true with $u_0 = u_1 = u_2 = \alpha$ for any positive scalar $\alpha$.

As in the case of Theorem 4.2.5, there are points that satisfy the Fritz John conditions trivially. If a feasible point $\bar{x}$ satisfies $\nabla f(\bar{x}) = 0$ or $\nabla g_i(\bar{x}) = 0$ for some $i \in I$, clearly we can let the corresponding Lagrangian multiplier be any positive number, set all the other multipliers equal to zero, and satisfy the conditions of Theorem 4.2.8. The Fritz John conditions of Theorem 4.2.8 also hold true trivially at each feasible point for problems having equality constraints if each equality constraint is replaced by two equivalent inequalities. Specifically, if $g(x) = 0$ is replaced by $g_1(x) \equiv g(x) \leq 0$ and $g_2(x) \equiv -g(x) \leq 0$, the Fritz John conditions are satisfied by taking $u_1 = u_2 = \alpha$ and setting all the other multipliers equal to zero, where $\alpha$ is any positive scalar.
The Fritz John and Karush–Kuhn–Tucker Optimality Conditions

Figure 4.8 Feasible Region in Example 4.2.11.

In fact, given any feasible solution \( \overline{x} \) to the problem of minimizing \( f(x) \) subject to \( x \in S \), we can add a redundant constraint to the problem to make \( \overline{x} \) a FJ point! Specifically, we can add the constraint \( \|x - \overline{x}\|^2 \geq 0 \), which holds true for all \( x \in \mathbb{R}^n \). In particular, this constraint is binding at \( \overline{x} \) and its gradient is also zero at \( \overline{x} \). Consequently, we obtain \( F_0 \cap G_0 = \emptyset \) at \( \overline{x} \) since \( G_0 = \emptyset \); so \( \overline{x} \) is a FJ point!

This leads us to consider two issues. The first pertains to a set of conditions under which we can claim local optimality for a FJ point, and this is addressed in Theorem 4.2.12. The second consideration leads to the Karush–Kuhn–Tucker necessary optimality conditions, and this is addressed subsequently.

4.2.12 Theorem (Fritz John Sufficient Conditions)

Let \( X \) be a nonempty open set in \( \mathbb{R}^n \), and let \( f: \mathbb{R}^n \to \mathbb{R} \) and \( g_i: \mathbb{R}^n \to \mathbb{R} \) for \( i = 1, \ldots, m \). Consider Problem P to minimize \( f(x) \) subject to \( x \in S \), and \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \). Let \( \overline{x} \) be a FJ solution and denote \( I = \{ i : g_i(\overline{x}) = 0 \} \). Define \( S \) as the relaxed feasible region for Problem P in which the nonbinding constraints are dropped.

a. If there exists an \( \epsilon \)-neighborhood \( N_\epsilon(\overline{x}) \), \( \epsilon > 0 \), such that \( f \) is pseudoconvex over \( N_\epsilon(\overline{x}) \cap S \), and \( g_i, i \in I \), are strictly pseudoconvex over \( N_\epsilon(\overline{x}) \cap S \), \( \overline{x} \) is a local minimum for Problem P.

b. If \( f \) is pseudoconvex at \( \overline{x} \), and if \( g_i, i \in I \), are both strictly pseudoconvex and quasiconvex at \( \overline{x} \), \( \overline{x} \) is a global optimal solution for Problem P. In particular, if these generalized convexity assumptions
hold true only by restricting the domain of $f$ to $N_{\varepsilon}(\bar{x})$ for some $\varepsilon > 0$, $\bar{x}$ is a local minimum for Problem P.

**Proof**

Suppose that the condition of Part a holds true. Since $\bar{x}$ is a FJ point, we have equivalently by Gordan’s theorem that $F_0 \cap G_0 = \emptyset$. By restricting attention to $S \cap N_{\varepsilon}(\bar{x})$, we have, by closely following the proof to the converse statement of Theorem 4.2.5, that $\bar{x}$ is a local minimum. This proves Part a.

Next, consider Part b. Again we have $F_0 \cap G_0 = \emptyset$. By restricting attention to $S$, we have $G_0 = D$ by Lemma 4.2.4; so we conclude that $F_0 \cap D = \emptyset$. Now let $x$ be any feasible solution to the relaxed constraint set $S$. [In the case when the generalized convexity assumptions hold true over $N_{\varepsilon}(\bar{x})$ alone, let $x \in S \cap N_{\varepsilon}(\bar{x})$.] Since $g_i(x) \leq g_i(\bar{x}) = 0$ for all $i \in I$, we have, by the quasiconvexity of $g_i$ at $\bar{x}$ for all $i \in I$, that

$$g_i[x + \lambda(x - \bar{x})] = g_i[\lambda x + (1 - \lambda)\bar{x}] \leq \max\{g_i(x), g_i(\bar{x})\} = g_i(\bar{x}) = 0$$

for all $0 \leq \lambda \leq 1$, and for each $i \in I$. This means that the direction $d = (x - \bar{x}) \in D$. Because $F_0 \cap D = \emptyset$, we must therefore have $\nabla f(\bar{x})^t d \geq 0$; that is, $\nabla f(\bar{x})^t (x - \bar{x}) \geq 0$. By the pseudoconvexity of $f$ at $\bar{x}$, this in turn implies that $f(x) \geq f(\bar{x})$. Hence, $\bar{x}$ is a global optimum over the relaxed set $S$ [or over $S \cap N_{\varepsilon}(\bar{x})$ in the second case]. And since it belongs to the original feasible region or to its intersection with $N_{\varepsilon}(\bar{x})$, it is a global (or a local in the second case) minimum for Problem P. This completes the proof.

We remark here that as is evident from the analysis thus far, several variations of the assumptions in Theorem 4.2.12 are possible. We encourage the reader to explore this in Exercise 4.22.

**Karush–Kuhn–Tucker Conditions**

We have observed above that a point $\bar{x}$ is a FJ point if and only if $F_0 \cap G_0 = \emptyset$. In particular, this condition holds true at any feasible solution $\bar{x}$ at which $G_0 = \emptyset$, regardless of the objective function. For example, if the feasible region has no interior in the immediate vicinity of $\bar{x}$, or if the gradient of some binding constraint (which might even be redundant) vanishes, $G_0 = \emptyset$. Generally speaking, by Gordan’s theorem, $G_0 = \emptyset$ if and only if the gradients of the binding constraints can be made to cancel out using nonnegative, nonzero linear combinations, and whenever this case occurs, $\bar{x}$ will be a FJ point. More disturbingly, it follows that FJ points can be nonoptimal even for the well-
The Fritz John and Karush–Kuhn–Tucker Optimality Conditions

behaved and important class of linear programming (LP) problems. Figure 4.9 illustrates this situation.

Motivated by this observation, we are led to the KKT conditions described next that encompass FJ points for which there exist Lagrange multipliers such that \( u_0 > 0 \) and hence force the objective function gradient to play a role in the optimality conditions. These conditions were derived independently by Karush [1939] and by Kuhn and Tucker [1951], and are precisely the FJ conditions with the added requirement that \( u_0 > 0 \). Note that when \( u_0 > 0 \), by scaling the dual feasibility conditions, if necessary, we can assume without loss of generality that \( u_0 = 1 \). Hence, in Example 4.2.9, taking \( u_0 = 1 \) in the FJ conditions, we obtain \( (u_0, u_1, u_2) = (1, 1/3, 2/3) \) as the Lagrange multipliers corresponding to the optimal solution. Moreover, in Figure 4.9, the only FJ point that is also a KKT point is the optimal solution \( \bar{x} \). In fact, as we shall see later, the KKT conditions are both necessary and sufficient for optimality for linear programming problems. Example 4.2.11 gives another illustration of a linear programming problem.

Also, note from the above discussion that if \( G_0 \neq \emptyset \) at a local minimum \( \bar{x} \), \( \bar{x} \) must indeed be a KKT point; that is, it must be a FJ point with \( u_0 > 0 \). This follows because by Gordan’s theorem, if \( G_0 \neq \emptyset \), no solution exists to FJ’s dual feasibility conditions with \( u_0 = 0 \). Hence, \( G_0 \neq \emptyset \) is a sufficient condition placed on the behavior of the constraints to ensure that a local minimum \( \bar{x} \) is a KKT point. Of course, it need not necessarily hold true whenever a local minimum \( \bar{x} \) turns out to be a KKT point, as in Figure 4.9, for example. Such a condition is known as a constraint qualification (CQ). Several conditions of this kind are

![Figure 4.9 FJ conditions are not sufficient for optimality for LP problems.](image-url)
discussed in more detail later and in Chapter 5. Note that the importance of constraint qualifications is to guarantee that by examining only KKT points, we do not lose out on local minima and hence, possibly, global optimal solutions. This can certainly occur, as is evident from Figure 4.7 of Example 4.2.10, where $u_0$ is necessarily zero in the FJ conditions for the optimal solution.

In Theorem 4.2.13, by imposing the constraint qualification that the gradient vectors of the binding constraints are linearly independent, we obtain the KKT conditions. Note that if the gradients of the binding constraints are linearly independent, certainly they cannot be canceled by using nonzero, nonnegative, linear combinations; and hence, this implies by Gordan’s theorem that $G_0 \neq \emptyset$. Therefore, the linear independence constraint qualification implies the constraint qualification that $G_0 \neq \emptyset$; and hence, as above, it implies that a local minimum $\bar{x}$ satisfies the KKT conditions. This is formalized below.

4.2.13 Theorem (Karush-Kuhn-Tucker Necessary Conditions)

Let $X$ be a nonempty open set in $\mathbb{R}^n$, and let $f: \mathbb{R}^n \to \mathbb{R}$ and $g_i: \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$. Consider the Problem P to minimize $f(x)$ subject to $x \in X$ and $g_i(x) \leq 0$ for $i = 1, \ldots, m$. Let $\bar{x}$ be a feasible solution, and denote $I = \{i : g_i(\bar{x}) = 0\}$.

Suppose that $f$ and $g_i$ for $i \in I$ are differentiable at $\bar{x}$ and that $g_i$ for $i \notin I$ are continuous at $\bar{x}$. Furthermore, suppose that $\nabla g_i(\bar{x})$ for $i \in I$ are linearly independent. If $\bar{x}$ solves Problem P locally, there exist scalars $u_i$ for $i \in I$ such that

$$
\nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) = 0
$$

$$
u_i \geq 0 \quad \text{for } i \in I.
$$

In addition to the above assumptions, if $g_i$ for each $i \notin I$ is also differentiable at $\bar{x}$, the foregoing conditions can be written in the following equivalent form:

$$
\nabla f(\bar{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) = 0
$$

$$
u_i = 0 \quad \text{for } i = 1, \ldots, m
$$

$$
u_i \geq 0 \quad \text{for } i = 1, \ldots, m.
$$

Proof

By Theorem 4.2.8 there exist scalars $u_0$ and $\hat{u}_i$ for $i \in I$, not all equal to zero, such that

$$
u_0 \nabla f(\bar{x}) + \sum_{i \in I} \hat{u}_i \nabla g_i(\bar{x}) = 0$$

$$u_0, \hat{u}_i \geq 0 \quad \text{for } i \in I.$$

(4.10)
The Fritz John and Karush-Kuhn-Tucker Optimality Conditions

Note that $u_0 > 0$ because (4.10) would contradict the assumption of linear independence of $\nabla g_i(\bar{x})$ for $i \in I$ if $u_0 = 0$. The first part of the theorem follows by letting $u_i = \hat{u}_i / u_0$ for each $i \in I$. The equivalent form of the necessary conditions follows by letting $u_i = 0$ for $i \not\in I$. This completes the proof.

As in the Fritz John conditions, the scalars $u_i$ are called the Lagrangian, or Lagrange, multipliers. The requirement that $\bar{x}$ be feasible to Problem P is called the primal feasibility (PF) condition, whereas the condition that $\nabla f(\bar{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) = 0, u_i \geq 0$ for $i = 1,..., m$, is referred to as the dual feasibility (DF) condition. The restriction $u_i g_i(\bar{x}) = 0$ for each $i = 1,..., m$ is called the complementary slackness (CS) condition. Together, these PF, DF, and CS conditions are called the Karush-Kuhn-Tucker conditions. Any point $\bar{x}$ for which there exist Lagrangian (or Lagrange) multipliers $\bar{u}$ such that $(\bar{x}, \bar{u})$ satisfies the KKT conditions is called a KKT point. Note that if the gradients $\nabla g_i(\bar{x})$, $i \in I$, are linearly independent, then by the DF and CS conditions, the associated Lagrange multipliers are determined uniquely at the KKT point $\bar{x}$.

The KKT conditions can alternatively be written in vector form as follows, in addition to the PF requirement:

$$\nabla f(\bar{x}) + \nabla g(\bar{x})^T u = 0$$
$$u^T g(\bar{x}) = 0$$
$$u \geq 0.$$

Here $\nabla g(\bar{x})^T$ is an $n \times m$ matrix whose $i$th column is $\nabla g_i(\bar{x})$ (it is the transpose of the Jacobian of $g$ at $\bar{x}$), and $u$ is an $m$-vector denoting the Lagrangian multipliers.

Now, consider Examples 4.2.9, 4.2.10, and 4.2.11. In Example 4.2.9, at $\bar{x} = (2,1)^T$, the reader may verify that $u_1 = 1/3, u_2 = 2/3,$ and $u_3 = u_4 = 0$ will satisfy the KKT conditions. Example 4.2.10 does not satisfy the assumptions of Theorem 4.2.13 at $\bar{x} = (1,0)^T$, since $\nabla g_1(\bar{x})$ and $\nabla g_2(\bar{x})$ are linearly dependent. In fact, in this example, we saw that $u_0$ is necessarily zero in the FJ conditions. In Example 4.2.11, $\bar{x} = (1,0)^T$ and $u_1 = u_2 = 1$ satisfy the KKT conditions.

4.2.14 Example (Linear Programming Problems)

Consider the linear programming Problem P: Minimize $\{c^T x: Ax = b, x \geq 0\}$, where $A$ is $m \times n$ and the other vectors are conformable. Writing the constraints as $-Ax \leq -b$, $Ax \leq b$, and $-x \leq 0$, and denoting the Lagrange multiplier vectors with respect to these three sets as $y^+$, $y^-$, and $v$, respectively, the KKT conditions are as follows:
PF: \[ Ax = b, \ x \geq 0 \]

DF: \[ -A'y^+ + A'y^- - v = -c, \ (y^+, y^-, v) \geq 0 \]

CS: \[ (b-Ax)^iy^+ = 0, \quad (Ax-b)^iy^- = 0, \quad -x^tv = 0. \]

Denoting \( y = y^+ - y^- \) as the difference of the two nonnegative variable vectors \( y^+ \) and \( y^- \), we can equivalently write the KKT conditions as follows, noting the use of the PF and DF conditions in simplifying the CS conditions:

PF: \[ Ax = b, \ x \geq 0 \]

DF: \[ A'y + v = c, \ v \geq 0, \quad (y \text{ unrestricted}) \]

CS: \[ x_jv_j = 0 \quad \text{for} \ j = 1, \ldots, n. \]

Hence, from Theorem 2.7.3 and its Corollary 3, observe that \( \bar{x} \) is KKT solution with associated Lagrange multipliers \((\bar{y}, \bar{v})\) if and only if \( \bar{x} \) and \( \bar{y} \) are, respectively, optimal to the primal and dual linear programs \( P \) and \( D \), where \( D: \)
Maximize \[ \{b'y : A'y \leq c\}. \]
In particular, observe that the DF restriction in the KKT conditions is precisely the feasibility condition to the dual \( D \); hence, the name. This example therefore establishes that for linear programming problems, the KKT conditions are both necessary and sufficient for optimality to the primal and dual problems.

**Geometric Interpretation of the Karush–Kuhn–Tucker Conditions: Connections with Linear Programming Approximations**

Note that any vector of the form \( \sum_{i \in I} u_i \nabla g_i(\bar{x}) \), where \( u_i \geq 0 \) for \( i \in I \), belongs to the cone spanned by the gradients of the binding constraints. The KKT dual feasibility conditions \( -\nabla f(\bar{x}) = \sum_{i \in I} u_i \nabla g_i(\bar{x}) \) and \( u_i \geq 0 \) for \( i \in I \) can then be interpreted as \( -\nabla f(\bar{x}) \), belonging to the cone spanned by the gradients to the binding constraints at a given feasible solution \( \bar{x} \).

Figure 4.10 illustrates this concept for two points \( x_1 \) and \( x_2 \). Note that \( -\nabla f(x_1) \) belongs to the cone spanned by the gradients of the binding constraints at \( x_1 \) and hence \( x_1 \) is a KKT point; that is, \( x_1 \) satisfies the KKT conditions. On the other hand, \( -\nabla f(x_2) \) lies outside the cone spanned by the gradients of the binding constraints at \( x_2 \) and thus contradicts the KKT conditions.

Similarly, in Figures 4.6 and 4.8, for \( \bar{x} = (2,1)' \) and \( \bar{x} = (1,0)' \), respectively, \( -\nabla f(\bar{x}) \) lies in the cone spanned by the gradients of the binding constraints at \( \bar{x} \). On the other hand, in Figure 4.7, for \( \bar{x} = (1,0)' \), \( -\nabla f(\bar{x}) \) lies outside the cone spanned by the gradients of the binding constraints at \( \bar{x} \).
The Fritz John and Karush–Kuhn–Tucker Optimality Conditions

193

Figure 4.10 Geometric illustration of the KKT conditions.

We now provide a key insight into the KKT conditions via linear programming duality and Farkas’s lemma as expounded in Theorem 2.7.3 and its Corollary 2. The following result asserts that a feasible solution \( \bar{x} \) is a KKT point if and only if it happens to be an optimal to the linear program obtained by replacing the objective and the constraints by their first-order approximations at \( \bar{x} \). (This is referred to as the first-order linear programming approximation to the problem at \( \bar{x} \).) Not only does this provide a useful conceptual characterization of KKT points and an insight into its value and interpretation, but it affords a useful construct in deriving algorithms that are designed to converge to a KKT solution.

4.2.15 Theorem (KKT Conditions and First-Order LP Approximations)

Let \( X \) be a nonempty open set in \( \mathbb{R}^n \), and let \( f : \mathbb{R}^n \to \mathbb{R} \) and \( g_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, m \) be differentiable functions. Consider Problem P, to minimize \( f(x) \) subject to \( x \in S = \{ x \in X : g_i(x) \leq 0, i = 1, \ldots, m \} \). Let \( \bar{x} \) be a feasible solution, and denote \( I = \{ i : g_i(\bar{x}) = 0 \} \). Define \( F_0 = \{ d : \nabla f(\bar{x})^t d < 0 \} \) and \( G'_0 = \{ d \neq 0 : \nabla g_i(\bar{x})^t d \leq 0, \text{ for each } i \in I \} \). Then \( \bar{x} \) is a KKT solution if and only if \( F_0 \cap G'_0 = \emptyset \). Furthermore, consider the first-order linear programming approximation to Problem P:
LP(\(\bar{x}\)) : Minimize \(\{f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) : g_i(\bar{x}) + \nabla g_i(\bar{x})^T (x - \bar{x}) \leq 0 \}
\text{for } i = 1,\ldots, m\).

Then, \(\bar{x}\) is a KKT solution if and only if \(\bar{x}\) solves LP(\(\bar{x}\)).

**Proof**

The feasible solution \(\bar{x}\) is a KKT point if and only if there exists a solution \((u_i, i \in I)\) to the system \(\sum_{i \in I} u_i \nabla g_i(\bar{x}) = -\nabla f(\bar{x})\) and \(u_i \geq 0\) for \(i \in I\). By Farkas's lemma (see, e.g., Corollary 2 to Theorem 2.7.3), this holds true if and only if there does not exist a solution to the system \(\nabla g_i(\bar{x})^T d \leq 0\) for \(i \in I\) and \(\nabla f(\bar{x})^T d < 0\). Hence, \(\bar{x}\) is a KKT point if and only if \(F_0 \cap G' = \emptyset\). Clearly, we also have that this holds true if and only if \(F_0 \cap G'_0 = \emptyset\).

Now consider the first-order linear programming approximation LP(\(\bar{x}\)) given in the theorem. The solution \(\bar{x}\) is obviously feasible to LP. Ignoring the constant terms in the objective function and writing LP(\(\bar{x}\)) in the form of Problem D of Theorem 2.7.3, we get that, equivalently, LP(\(\bar{x}\)): Maximize \(\{-\nabla f(\bar{x})^T x : \nabla g_i(\bar{x})^T x \leq \nabla g_i(\bar{x})^T \bar{x} - g_i(\bar{x})\text{ for } i = 1,\ldots, m\}\). The dual to this problem, denoted DLP(\(\bar{x}\)), is to

\[
\begin{align*}
\text{Minimize} & \quad \sum_{i=1}^{m} u_i [\nabla g_i(\bar{x})^T \bar{x} - g_i(\bar{x})] \\
\text{subject to} & \quad \sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) = -\nabla f(\bar{x}), \quad u_i \geq 0 \quad \text{for } i = 1,\ldots, m.
\end{align*}
\]

Hence, by Corollary 3 to Theorem 2.7.3, we deduce that \(\bar{x}\) is an optimal solution to LP(\(\bar{x}\)) if and only if there exists a solution \(\bar{u}\) feasible to DLP(\(\bar{x}\)) that also satisfies the complementary slackness condition \(\bar{u}_i [\nabla g_i(\bar{x})^T \bar{x} - \nabla g_i(\bar{x})^T \bar{x} + g_i(\bar{x})] = \bar{u}_i g_i(\bar{x}) = 0\) for \(i = 1,\ldots, m\). But this is precisely the KKT conditions. Hence, \(\bar{x}\) is optimal to LP(\(\bar{x}\)) if and only if \(\bar{x}\) is a KKT solution for \(P\), and this completes the proof.

To illustrate, observe that in Figure 4.6 of Example 4.2.9, if we replace \(g_i(x) \leq 0\) by its tangential first-order approximation at the point (2, 1) and the objective function by the linear objective of minimizing \(\nabla f(\bar{x})^T x\), the given point (2, 1) is optimal to the resulting linear programming problem and hence is a KKT solution. On the other hand, in Figure 4.7 of Example 4.2.10, the feasible region for the linear programming approximation at \(\bar{x} = (1, 0)^T\) is the entire \(x_1\)-axis. Clearly, then, the point (1, 0) is not optimal to the underlying linear program LP(\(x\)) of minimizing \(\nabla f(\bar{x})^T x\) over this region, and thus the point (1, 0) is not a KKT point. Hence, the KKT conditions, being oblivious to the nonlinear
behavior of the constraint \( g_1(x) \leq 0 \) about the point \( \bar{x} \) other than its first-order approximation, fail to recognize the optimality of this solution for the original nonlinear problem.

Theorem 4.2.16 shows that under convexity assumptions, the KKT conditions are also sufficient for (local) optimality.

**4.2.16 Theorem (Karush–Kuhn–Tucker Sufficient Conditions)**

Let \( X \) be a nonempty open set in \( \mathbb{R}^n \), and let \( f: \mathbb{R}^n \to \mathbb{R} \) and \( g_i: \mathbb{R}^n \to \mathbb{R} \) for \( i = 1, \ldots, m \). Consider Problem P, to minimize \( f(x) \) subject to \( x \in X \) and \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \). Let \( \bar{x} \) be a KKT solution, and denote \( I = \{ i : g_i(\bar{x}) = 0 \} \). Define \( S \) as the relaxed feasible region for Problem P in which the constraints that are not binding at \( \bar{x} \) are dropped. Then:

a. If there exists an \( \varepsilon \)-neighborhood \( N_{\varepsilon}(\bar{x}) \) about \( \bar{x} \), \( \varepsilon > 0 \) such that \( f \) is pseudoconvex over \( N_{\varepsilon}(\bar{x}) \cap S \) and \( g_i, i \in I, \) are differentiable at \( \bar{x} \) and are quasiconvex over \( N_{\varepsilon}(\bar{x}) \cap S \), \( \bar{x} \) is a local minimum for Problem P.

b. If \( f \) is pseudoconvex at \( \bar{x} \), and if \( g_i, i \in I, \) are differentiable and quasiconvex at \( \bar{x} \), \( \bar{x} \) is a global optimal solution to Problem P. In particular, if this assumption holds true with the domain of the feasible restriction to \( N_{\varepsilon}(\bar{x}) \), for some \( \varepsilon > 0 \), \( \bar{x} \) is a local minimum for P.

**Proof**

First, consider Part a. Since \( \bar{x} \) is a KKT point, we have, equivalently, by Theorem 4.2.15 that \( F_0 \cap G_0 = \emptyset \). From (4.7) this means that \( F_0 \cap D = \emptyset \). Since \( g_i, i \in I, \) are quasiconvex over \( N_{\varepsilon}(\bar{x}) \cap S \), we have that \( N_{\varepsilon}(\bar{x}) \cap S \) is a convex set. By restricting attention to \( N_{\varepsilon}(\bar{x}) \cap S \), we therefore have the condition of the converse statement of Theorem 4.2.2 holding true; so \( \bar{x} \) is a minimum over \( N_{\varepsilon}(\bar{x}) \cap S \). Hence, \( \bar{x} \) is a local minimum for the more restricted original Problem P. This proves Part a.

Next, consider Part b. Let \( x \) be any feasible solution to Problem P. [In case the generalized convexity definitions are restricted to \( N_{\varepsilon}(\bar{x}) \), let \( x \) be any feasible solution to P that lies within \( N_{\varepsilon}(\bar{x}) \).] Then, for \( i \in I, g_i(x) \leq g_i(\bar{x}), \) since \( g_i(x) \leq 0 \) and \( g_i(\bar{x}) = 0 \). By the quasiconvexity of \( g_i \) at \( \bar{x} \), it follows that

\[
g_i[\bar{x} + \lambda(x - \bar{x})] = g_i[\lambda x + (1 - \lambda)\bar{x}] \leq \max\{g_i(x), g_i(\bar{x})\} = g_i(\bar{x})
\]

for all \( \lambda \in (0, 1) \). This implies that \( g_i \) does not increase by moving from \( \bar{x} \) along the direction \( x - \bar{x} \). Thus, by Theorem 4.1.2 we must have \( \nabla g_i(\bar{x})'(x - \bar{x}) \leq 0. \)
Multiplying this by the Lagrange multiplier $u_i$ corresponding to the KKT point $\bar{x}$, and summing over $I$, we get $[\sum_{i \in I} u_i \nabla g_i(\bar{x})]'(x - \bar{x}) \leq 0$. But since $\nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) = 0$, it follows that $\nabla f(\bar{x})'(x - \bar{x}) \geq 0$. Then, by the pseudoconvexity of $f$ at $\bar{x}$, we must have $f(x) \geq f(\bar{x})$, and the proof is complete.

Needless to say, if $f$ and $g_i$ are convex at $\bar{x}$, and hence both pseudoconvex and quasiconvex at $\bar{x}$, the KKT conditions are sufficient. Also, if convexity at a point is replaced by the stronger requirement of global convexity, the KKT conditions are also sufficient for global optimality. (We ask the reader to explore other variations of this result in Exercises 4.22 and 4.50.)

There is one important point to note in regard to KKT conditions that is often a source of error. Namely, despite the usually well-behaved nature of convex programming problems and the sufficiency of KKT conditions under convexity assumptions, the KKT conditions are not necessary for optimality for convex programming problems. Figure 4.11 illustrates this situation for the convex programming problem given:

Minimize $x_1$
subject to $(x_1 - 1)^2 + (x_2 - 1)^2 \leq 1$
$(x_1 - 1)^2 + (x_2 + 1)^2 \leq 1$.

The only feasible solution $\bar{x} = (1, 0)'$ is naturally optimal. However, this is not a KKT point. Note in connection with Theorem 4.2.15 that the first-order linear programming approximation at $\bar{x}$ is unbounded. However, as we shall see in Chapter 5, if there exists an interior point feasible solution to the set of constraints that are binding at an optimum $\bar{x}$ to a convex programming problem, $\bar{x}$ is indeed a KKT point and is therefore captured by the KKT conditions.

Figure 4.11 KKT conditions are not necessary for convex programming problems.
4.3 Problems Having Inequality and Equality Constraints

In this section we generalize the optimality conditions of the preceding section to handle inequality constraints as well as equality constraints. Consider the following nonlinear programming Problem P:

Minimize \( f(x) \)
subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \)
\( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \)
\( x \in X. \)

As a natural extension of Theorem 4.2.5, in Theorem 4.3.1 we show that if \( \bar{x} \) is a local optimal solution to Problem P, either the gradients of the equality constraints are linearly dependent at \( \bar{x} \), or else,
\[ F_0 \cap G_0 \cap H_0 = \emptyset, \]
where \( H_0 = \{ d : \nabla h_i(\bar{x})'d = 0 \text{ for } i = 1, \ldots, \ell \}. \) A reader with only a casual interest in the derivation of optimality conditions may skip the proof of Theorem 4.3.1, since it involves the more advanced concepts of solving a system of differential equations.

4.3.1 Theorem

Let \( X \) be a nonempty open set in \( \mathbb{R}^n \). Let \( f: \mathbb{R}^n \to \mathbb{R}, g_i: \mathbb{R}^n \to \mathbb{R} \) for \( i = 1, \ldots, m \), and \( h_i: \mathbb{R}^n \to \mathbb{R} \) for \( i = 1, \ldots, \ell \). Consider the Problem P given below:

Minimize \( f(x) \)
subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \)
\( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \)
\( x \in X. \)

Suppose that \( \bar{x} \) is a local optimal solution, and let \( I = \{ i: g_i(\bar{x}) = 0 \} \). Furthermore, suppose that each \( g_i \) for \( i \notin I \) is continuous at \( \bar{x} \), that \( f \) and \( g_i \) for \( i \in I \) are differentiable at \( \bar{x} \) and that each \( h_i \) for \( i = 1, \ldots, \ell \) is continuously differentiable at \( \bar{x} \). If \( \nabla h_i(\bar{x}) \) for \( i = 1, \ldots, \ell \) are linearly independent, \( F_0 \cap G_0 \cap H_0 = \emptyset \), where
\[ F_0 = \{ d : \nabla f(\bar{x})'d < 0 \}, \]
\[ G_0 = \{ d : \nabla g_i(\bar{x})'d < 0 \text{ for } i \in I \}, \]
\[ H_0 = \{ d : \nabla h_i(\bar{x})'d = 0 \text{ for } i = 1, \ldots, \ell \}. \]

Conversely, suppose that \( F_0 \cap G_0 \cap H_0 = \emptyset \). If \( f \) is pseudoconvex at \( \bar{x} \), \( g_i \) for \( i \in I \) are strictly pseudoconvex over some \( \epsilon \)-neighborhood of \( \bar{x} \); and if \( h_i \) for \( i = 1, \ldots, \ell \) are affine, \( \bar{x} \) is a local optimal solution.
Proof

Consider the first part of the theorem. By contradiction, suppose that there exists a vector $y \in F_0 \cap G_0 \cap H_0$, that is, $\nabla f(\bar{x})^T y < 0$, $\nabla g_i(\bar{x})^T y < 0$ for each $i \in I$, and $\nabla h_i(\bar{x})y = 0$, where $\nabla h_i(\bar{x})$ is the $\ell \times n$ Jacobian matrix whose $i$th row is $\nabla h_i(\bar{x})^T$. Let us now construct a feasible arc from $\bar{x}$ obtained by projecting points along $y$ from $\bar{x}$ onto the equality-constraint surface. For $\lambda \geq 0$, define $\alpha : R \to R^n$ by the following differential equation and boundary condition:

$$\frac{d\alpha(\lambda)}{d\lambda} = P(\lambda)y \quad \text{and} \quad \alpha(0) = \bar{x}, \quad (4.11)$$

where $P(\lambda)$ is the matrix that projects any vector in the null space of $\nabla h[\alpha(\lambda)]$. For $\lambda$ sufficiently small, (4.11) is well defined and solvable because $\nabla h(\bar{x})$ has full rank and $h$ is continuously differentiable at $\bar{x}$, so that $P$ is continuous in $\lambda$. Obviously, $\alpha(\lambda) \to \bar{x}$ as $\lambda \to 0^+$.

We now show that for $\lambda > 0$ and sufficiently small, $\alpha(\lambda)$ is feasible and $f[\alpha(\lambda)] < f(\bar{x})$, thus contradicting local optimality of $\bar{x}$. By the chain rule of differentiation and from (4.11), we get

$$\frac{d}{d\lambda} g_i[\alpha(\lambda)] = \nabla g_i[\alpha(\lambda)]^T P(\lambda)y \quad (4.12)$$

for each $i \in I$. In particular, $y$ is in the null space of $\nabla h(\bar{x})$, so for $\lambda = 0$, we have $P(0)y = y$. Hence, from (4.12) and the fact that $\nabla g_i(\bar{x})^T y < 0$, we get

$$\frac{d}{d\lambda} g_i[\alpha(0)] = \nabla g_i(\bar{x})^T y < 0 \quad (4.13)$$

for $i \in I$. This implies further that $g_i[\alpha(\lambda)] < 0$ for $\lambda > 0$ and sufficiently small. For $i \not\in I$, $g_i(\bar{x}) < 0$, and $g_i$ is continuous at $\bar{x}$, and thus $g_i[\alpha(\lambda)] < 0$ for $\lambda$ sufficiently small. Also, since $X$ is open, $\alpha(\lambda) \in X$ for $\lambda$ sufficiently small. To show feasibility of $\alpha(\lambda)$, we only need to show that $h_i[\alpha(\lambda)] = 0$ for $\lambda$ sufficiently small. By the mean value theorem, we have

$$h_i[\alpha(\lambda)] = h_i[\alpha(0)] + \lambda \frac{d}{d\lambda} h_i[\alpha(\mu)]$$

$$= \lambda \frac{d}{d\lambda} h_i[\alpha(\mu)] \quad (4.14)$$

for some $\mu \in (0, \lambda)$. But by the chain rule of differentiation and similar to (4.12), we get
By construction, \( P(\mu)y \) is in the null space of \( \nabla h_i[\alpha(\mu)] \) and, hence, from the above equation, we get \( (d/d\lambda)h_i[\alpha(\mu)] = 0 \). Substituting in (4.14), it follows that \( h_i[\alpha(\lambda)] = 0 \). Since this is true for each \( i \), it follows that \( \alpha(\lambda) \) is a feasible solution to Problem P for each \( \lambda > 0 \) and is sufficiently small. By an argument similar to that leading to (4.13), we get

\[
\frac{d}{d\lambda} f[\alpha(0)] = \nabla f(\bar{x})^T y < 0
\]

and hence \( f[\alpha(\lambda)] < f(\bar{x}) \) for \( \lambda > 0 \) and sufficiently small. This contradicts the local optimality of \( \bar{x} \). Hence, \( F_0 \cap G_0 \cap H_0 = \emptyset \).

Conversely, suppose that \( F_0 \cap G_0 \cap H_0 = \emptyset \) and that the assumptions of the converse statement to the theorem hold true. Since \( h_i, i = 1,\ldots,\ell \), are affine, we have that \( d \) is a feasible direction for the equality constraints if and only if \( d \in H_0 \). Using Lemma 4.2.4, it is readily verified that since \( g_i, i \in I \), are strictly pseudoconvex over \( N_\varepsilon(\bar{x}) \) for some \( \varepsilon > 0 \), we have that \( D = G_0 \cap H_0 \), where \( D \) is the set of feasible directions at \( \bar{x} \) defined for the set \( S = \{ x : g_i(x) \leq 0 \text{ for } i \in I, \quad h_i(x) = 0 \text{ for } i = 1,\ldots,\ell \} \). Hence, we have that \( F_0 \cap D = \emptyset \). Moreover, by our assumptions, we know that \( S \cap N_\varepsilon(\bar{x}) \) is a convex set and that \( f \) is pseudoconvex at \( \bar{x} \). Hence, by the converse to Theorem 4.2.2, \( \bar{x} \) is a minimum over \( S \cap N_\varepsilon(\bar{x}) \). Therefore, \( \bar{x} \) is a local minimum for the more restricted original problem as well, and this completes the proof.

**Fritz John Conditions**

We now express the geometric optimality condition \( F_0 \cap G_0 \cap H_0 = \emptyset \) in a more usable algebraic form. This is done in Theorem 4.3.2, which is a generalization of the Fritz John conditions of Theorem 4.2.6.

**4.3.2 Theorem (Fritz John Necessary Conditions)**

Let \( X \) be a nonempty open set in \( \mathbb{R}^n \), and let \( f: \mathbb{R}^n \to \mathbb{R} \), \( g_i: \mathbb{R}^n \to \mathbb{R} \) for \( i = 1,\ldots, m \), and \( h_i: \mathbb{R}^n \to \mathbb{R} \) for \( i = 1,\ldots, \ell \). Consider Problem P defined below:

Minimize \( f(x) \)

subject to \( g_i(x) \leq 0 \) for \( i = 1,\ldots, m \)

\( h_i(x) = 0 \) for \( i = 1,\ldots, \ell \)

\( x \in X \).
Let \( \bar{x} \) be a feasible solution, and let \( I = \{ i : g_i(\bar{x}) = 0 \} \). Furthermore, suppose that \( g_i \) for each \( i \in I \) is continuous at \( \bar{x} \), that \( f_i \) and \( g_i \) for \( i \in I \) are differentiable at \( \bar{x} \) and that \( h_i \) for each \( i = 1, \ldots, \ell \) is continuously differentiable at \( \bar{x} \). If \( \bar{x} \) solves Problem P locally, there exist scalars \( u_0, u_i \) for \( i \in I \), and \( v_i \) for \( i = 1, \ldots, \ell \) such that

\[
\begin{align*}
\sum_{i \in I} u_i \nabla f_i(\bar{x}) + \sum_{i = 1}^\ell v_i \nabla h_i(\bar{x}) &= 0 \\
u_0, u_i &\geq 0 \quad \text{for } i \in I \\
(u_0, u_I, v) &\neq (0, 0, 0),
\end{align*}
\]

where \( u_f \) is the vector whose components are \( u_i \) for \( i \in I \) and \( v = (v_1, \ldots, v_\ell)' \). Furthermore, if each \( g_i \) for \( i \in I \) is also differentiable at \( \bar{x} \), the Fritz John conditions can be written in the following equivalent form, where \( u = (u_1, \ldots, u_m)' \) and \( v = (v_1, \ldots, v_\ell)' \):

\[
\begin{align*}
\sum_{i = 1}^m u_i \nabla g_i(\bar{x}) + \sum_{i = 1}^\ell v_i \nabla h_i(\bar{x}) &= 0 \\
u_i g_i(\bar{x}) &= 0 \quad \text{for } i = 1, \ldots, m \\
u_0, u_i &\geq 0 \quad \text{for } i = 1, \ldots, m \\
(u_0, u, v) &\neq (0, 0, 0).
\end{align*}
\]

**Proof**

If \( \nabla h_i(\bar{x}) \) for \( i = 1, \ldots, \ell \) are linearly dependent, we can find scalars \( v_1, \ldots, v_\ell \), not all zero, such that \( \sum_{i = 1}^\ell v_i \nabla h_i(\bar{x}) = 0 \). Letting \( u_0, u_i \) for \( i \in I \) be equal to zero, the conditions of the first part of the theorem hold trivially.

Now, suppose that \( \nabla h_i(\bar{x}) \) for \( i = 1, \ldots, \ell \) are linearly independent. Let \( A_1 \) be the matrix whose rows are \( \nabla f(\bar{x})' \) and \( \nabla g_i(\bar{x})' \) for \( i \in I \), and let \( A_2 \) be the matrix whose rows are \( \nabla h_i(\bar{x})' \) for \( i = 1, \ldots, \ell \). Then, from Theorem 4.3.1, the local optimality of \( \bar{x} \) implies that the system

\[
A_1 d < 0, \quad A_2 d = 0
\]

is inconsistent. Now, consider the following two sets:

\[
S_1 = \{(z_1, z_2) : z_1 = A_1 d, \ z_2 = A_2 d\} \\
S_2 = \{(z_1, z_2) : z_1 < 0, \ z_2 = 0\}.
\]
The Fritz John and Karush–Kuhn–Tucker Optimality Conditions

Note that $S_1$ and $S_2$ are nonempty convex sets such that $S_1 \cap S_2 = \emptyset$. Then by Theorem 2.4.8 there exists a nonzero vector $p' = (p'_1, p'_2)$ such that

$$p'_1 A_1 d + p'_2 A_2 d \geq p'_1 z_1 + p'_2 z_2 \quad \text{for each } d \in \mathbb{R}^n \text{ and } (z_1, z_2) \in \text{cl} S_2.$$

Letting $z_2 = 0$ and since each component of $z_1$ can be made an arbitrarily large negative number, it follows that $p_1 \geq 0$. Also, letting $(z_1, z_2) = (0, 0)$, we must have $(p'_1 A_1 + p'_2 A_2) d \geq 0$ for each $d \in \mathbb{R}^n$. Letting $d = -(A'_1 p_1 + A'_2 p_2)$, it follows that $-\left\| A'_1 p_1 + A'_2 p_2 \right\|^2 \geq 0$, and thus $A'_1 p_1 + A'_2 p_2 = 0$.

To summarize, we have shown that there exists a nonzero vector $p' = (p'_1, p'_2)$ with $p_1 \geq 0$ such that $A'_1 p_1 + A'_2 p_2 = 0$. Denoting the components of $p_1$ by $u_0$ and $u_i$ for $i \in I$, and letting $p_2 = v$, the first result follows. The equivalent form of the necessary conditions is readily obtained by letting $u_i = 0$ for $i \notin I$, and the proof is complete.

The reader may note that the Lagrangian multiplier $v_i$ associated with the $i$th equality constraint is unrestricted in sign. Note also that these conditions are not equivalently obtained by writing each equality as two associated inequalities and then applying the FJ conditions for the inequality-constrained case. The FJ conditions can also be written in vector notation as follows:

$$u_0 \nabla f'(\bar{x}) + \nabla g(\bar{x})' u + \nabla h(\bar{x})' v = 0$$

$$u' g(\bar{x}) = 0$$

$$(u_0, u) \geq (0, 0)$$

$$(u_0, u, v) \neq (0, 0, 0).$$

Here $\nabla g(\bar{x})$ is an $m \times n$ Jacobian matrix whose $i$th row is $\nabla g_i(\bar{x})'$, and $\nabla h(\bar{x})$ is an $\ell \times n$ Jacobian matrix whose $i$th row is $\nabla h_i(\bar{x})'$. Also, $u$ and $v$ are, respectively, an $m$-vector and an $\ell$-vector, denoting the Lagrangian multipliers associated with the inequality and equality constraints.

### 4.3.3 Example

Minimize $x_1^2 + x_2^2$

subject to $x_1^2 + x_2^2 \leq 5$

$$-x_1 \leq 0$$

$$-x_2 \leq 0$$

$$x_1 + 2x_2 = 4.$$
Here, we have only one equality constraint. We verify below that the Fritz John conditions hold true at the optimal point \( \mathbf{x} = (4/5, 8/5) \). First, note that there are no binding inequality constraints at \( \mathbf{x} \); that is, \( I = \emptyset \). Hence, the multipliers associated with the inequality constraints are equal to zero. Note that

\[
\nabla f(\mathbf{x}) = (8/5, 16/5) \quad \text{and} \quad \nabla h_1(\mathbf{x}) = (1, 2).
\]

Thus,

\[
u_0 \begin{pmatrix} 8 \\ 5 \\ 16 \\ 5 \end{pmatrix} + v_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}
\]

is satisfied, for example, by \( u_0 = 5 \) and \( v_1 = -8 \).

4.3.4 Example

Minimize \((x_1 - 3)^2 + (x_2 - 2)^2\)

subject to \(x_1^2 + x_2^2 \leq 5\)

\(-x_1 \leq 0\)

\(-x_2 \leq 0\)

\(x_1 + 2x_2 = 4\).

This example is the same as Example 4.2.9, with the inequality constraint \(x_1 + 2x_2 \leq 4\) replaced by \(x_1 + 2x_2 = 4\). At the optimal point \( \mathbf{x} = (2, 1) \), we have only one inequality constraint \(x_1^2 + x_2^2 \leq 5\) binding. The Fritz John condition

\[
u_0 \begin{pmatrix} -2 \\ -2 \end{pmatrix} + u_1 \begin{pmatrix} 4 \\ 2 \end{pmatrix} + v_1 \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}
\]

is satisfied, for example, by \( u_0 = 3 \), \( u_1 = 1 \), and \( v_1 = 2 \).

4.3.5 Example

Minimize \(-x_1\)

subject to \(x_2 - (1 - x_1)^3 = 0\)

\(-x_2 - (1 - x_1)^3 = 0\).

As shown in Figure 4.12, this problem has only one feasible point, namely, \( \mathbf{x} = (1, 0) \). At this point, we have

\[
\nabla f(\mathbf{x}) = (-1, 0) \quad \text{and} \quad \nabla h_1(\mathbf{x}) = (0, 1) \quad \text{and} \quad \nabla h_2(\mathbf{x}) = (0, -1).
\]
The Fritz John and Karush–Kuhn–Tucker Optimality Conditions

Figure 4.12 Setup for Example 4.3.5.

The condition

\[ u_0 \begin{pmatrix} -1 \\ 0 \end{pmatrix} + v_1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \]

holds true only if \( u_0 = 0 \) and \( v_1 = v_2 = \alpha \), where \( \alpha \) is any scalar. Thus, the Fritz John necessary conditions are met at the point \( \bar{x} \).

Similar to Theorem 4.2.12, we now provide a set of sufficient conditions that enable us to guarantee that a FJ point is a local minimum. Again, as with Theorem 4.2.12, several variations of such sufficient conditions are possible. We state the result below using one such condition motivated by the converse to Theorem 4.3.1, and ask the reader to explore other conditions in Exercise 4.22.

4.3.6 Theorem (Fritz John Sufficient Conditions)

Let \( X \) be a nonempty open set in \( \mathbb{R}^n \), and let \( f: \mathbb{R}^n \rightarrow \mathbb{R} \), \( g_i: \mathbb{R}^n \rightarrow \mathbb{R} \), \( i = 1, \ldots, m \), and \( h_i: \mathbb{R}^n \rightarrow \mathbb{R} \), \( i = 1, \ldots, \ell \). Consider the Problem P given below:

Minimize \( f(x) \)
subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \)
\( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \)
\( x \in X \).
Let $\bar{x}$ be a FJ solution and denote $I = \{i : g_i(\bar{x}) = 0\}$. Define $S = \{x : g_i(x) \leq 0$ for $i \in I$, $h_i(\bar{x}) = 0$, for $i = 1, \ldots, \ell\}$. If $h_i$ for $i = 1, \ldots, \ell$ are affine and $\nabla h_i(\bar{x})$, $i = 1, \ldots, \ell$, are linearly independent, and if there exists an $\varepsilon$-neighborhood $N_\varepsilon(\bar{x})$ of $\bar{x}$, $\varepsilon > 0$ such that $f$ is pseudoconvex on $S \cap N_\varepsilon(\bar{x})$, and $g_i$ for $i \in I$ are strictly pseudoconvex over $S \cap N_\varepsilon(\bar{x})$, $\bar{x}$ is a local minimum for Problem P.

**Proof**

Let us first show that $F_0 \cap G_0 \cap H_0 = \emptyset$, where these sets are as defined in Theorem 4.3.1. On the contrary, suppose that there exists a solution $d \in F_0 \cap G_0 \cap H_0$. Then, by taking the inner product of the dual feasibility condition $u_0 \nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) + \sum_{i = 1}^\ell v_i \nabla h_i(\bar{x}) = 0$ with $d$, we obtain $u_0 \nabla f(\bar{x})'d + \sum_{i \in I} u_i \nabla g_i(\bar{x})'d = 0$, since $d \in H_0$. But $d \in F_0 \cap G_0$ and $(u_0, u_i$ for $i \in I) \geq 0$ implies that $(u_0, u_i$ for $i \in I) = (0, 0)$. Since $\bar{x}$ is a FJ point, we therefore must have a solution to the system $\sum_{i = 1}^\ell v_i \nabla h_i(\bar{x}) = 0$, $\nabla h_i(\bar{x}) = 0$, which contradicts the linear independence of $\nabla h_i(\bar{x})$ for $i = 1, \ldots, \ell$. Hence, $F_0 \cap G_0 \cap H_0 = \emptyset$.

Now, closely following the proof to the converse statement of Theorem 4.3.1, and restricting attention to $S \cap N_\varepsilon(\bar{x})$, we can conclude that $\bar{x}$ is a local minimum for P. This completes the proof.

**Karush-Kuhn-Tucker Conditions**

In the Fritz John conditions, the Lagrangian multiplier associated with the objective function is not necessarily positive. Under further assumptions on the constraint set, we can claim that at any local minimum, there exists a set of Lagrange multipliers for which $u_0$ is positive. In Theorem 4.3.7, we obtain a generalization of the KKT necessary optimality conditions of Theorem 4.2.13. This is done by imposing a qualification on the gradients of the equality and binding inequality constraints that ensure that $u_0 > 0$ necessarily holds true in the Fritz John conditions. Other qualifications on the constraints to ensure the existence of $u_0 > 0$ in the FJ conditions at a local minimum are discussed in Chapter 5.

**4.3.7 Theorem (Karush-Kuhn-Tucker Necessary Conditions)**

Let $X$ be a nonempty open set in $\mathbb{R}^n$, and let $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$, and $h_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, \ell$. Consider the Problem P given below:
The Fritz John and Karush–Kuhn–Tucker Optimality Conditions

Minimize \( f(x) \)
subject to \( g_i(x) \leq 0 \) for \( i = 1,\ldots, m \)
\( h_i(x) = 0 \) for \( i = 1,\ldots, \ell \)
\( x \in X. \)

Let \( \bar{x} \) be a feasible solution, and let \( I = \{ i : g_i(\bar{x}) = 0 \} \). Suppose that \( f \) and \( g_i \) for \( i \in I \) are differentiable at \( \bar{x} \), that each \( g_i \) for \( i \notin I \) is continuous at \( \bar{x} \), and that each \( h_i \) for \( i = 1,\ldots, \ell \) is continuously differentiable at \( \bar{x} \). Further, suppose that \( \nabla g_i(\bar{x}) \) for \( i \in I \) and \( \nabla h_i(\bar{x}) \) for \( i = 1,\ldots, \ell \) are linearly independent. (Such an \( \bar{x} \) is sometimes called regular.) If \( \bar{x} \) solves Problem P locally, there exist unique scalars \( u_i \) for \( i \in I \) and \( v_j \) for \( i = 1,\ldots, \ell \) such that

\[
\nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) + \sum_{i = 1}^\ell v_j \nabla h_i(\bar{x}) = 0
\]

\[
u_i \geq 0 \quad \text{for } i \in I.
\]

In addition to the above assumptions, if each \( g_i \) for \( i \notin I \) is also differentiable at \( \bar{x} \), the KKT conditions can be written in the following equivalent form:

\[
\nabla f(\bar{x}) + \sum_{i = 1}^m u_i \nabla g_i(\bar{x}) + \sum_{i = 1}^\ell v_j \nabla h_i(\bar{x}) = 0
\]

\[
u_i g_i(\bar{x}) = 0 \quad \text{for } i = 1,\ldots, m
\]

\[
u_i \geq 0 \quad \text{for } i = 1,\ldots, m.
\]

**Proof**

By Theorem 4.3.2 there exist scalars \( u_0 \) and \( \hat{u}_i \) for \( i \in I \), and \( \hat{v}_i \) for \( i = 1,\ldots, \ell \), not all zero, such that

\[
u_0 \nabla f(\bar{x}) + \sum_{i \in I} \hat{u}_i \nabla g_i(\bar{x}) + \sum_{i = 1}^\ell \hat{v}_i \nabla h_i(\bar{x}) = 0 \quad (4.15)
\]

\[
u_0, \hat{u}_i \geq 0 \quad \text{for } i \in I.
\]

Note that \( u_0 > 0 \), because if \( u_0 = 0 \), (4.15) would contradict the assumption of linear independence of \( \nabla g_i(\bar{x}) \) for \( i \in I \) and \( \nabla h_i(\bar{x}) \) for \( i = 1,\ldots, \ell \). The first result then follows by letting \( u_i = \hat{u}_i/u_0 \) for \( i \in I \), and \( v_i = \hat{v}_i/u_0 \) for \( i = 1,\ldots, \ell \), and noting that the linear independence assumption implies the uniqueness of these Lagrangian multipliers. The equivalent form of the necessary conditions follows by letting \( u_i = 0 \) for \( i \notin I \). This completes the proof.
Note that the KKT conditions of Theorem 4.3.7 can be written in vector form as follows:

\[ \nabla f(\bar{x}) + \nabla g(\bar{x})' u + \nabla h(\bar{x})' v = 0 \]
\[ u' g(\bar{x}) = 0 \]
\[ \bar{u} \geq 0. \]

Here \( \nabla g(\bar{x}) \) is an \( m \times n \) Jacobian matrix and \( \nabla h(\bar{x}) \) is an \( \ell \times n \) Jacobian matrix whose \( i \)th rows, respectively, are \( \nabla g_i(\bar{x})' \) and \( \nabla h_i(\bar{x})' \). The vectors \( u \) and \( v \) are the Lagrangian multiplier vectors.

The reader might have observed that the KKT conditions of Theorem 4.3.6 are precisely the KKT conditions of the inequality case given in Theorem 4.2.13 when each equality constraint \( h_i(x) = 0 \) is replaced by the two equivalent inequalities \( h_i(x) \leq 0 \) and \(- h_i(x) \leq 0\), for \( i = 1, ..., \ell \). Denoting \( v_i^+ \) and \( v_i^- \) as the nonnegative Lagrangian multipliers associated with the latter two inequalities and using the KKT conditions for the inequality case produces the KKT conditions of Theorem 4.3.7 upon replacing the difference \( v_i^+ - v_i^- \) of two nonnegative variables by the unrestricted variable \( v_i \) for each \( i = 1, ..., \ell \). In fact, writing the equalities as equivalent inequalities, the sets \( G'_0 \) and \( G' \) defined in Theorem 4.2.15 become, respectively, \( G'_0 \cap H_0 \) and \( G' \cap H_0 \). Theorem 4.2.15 then asserts that for Problem P of the present section,

\[ \bar{x} \text{ is a KKT solution } \iff F_0 \cap G'_0 \cap H_0 = \emptyset \iff F_0 \cap G' \cap H_0 = \emptyset. \] (4.16)

Moreover, this happens if and only if \( \bar{x} \) solves the first-order linear programming approximation \( LP(\bar{x}) \) at the point \( \bar{x} \) given by

\[
LP(\bar{x}) : \text{Minimize}\{ f(\bar{x}) + \nabla f(\bar{x})' (x - \bar{x}) : g_i(\bar{x}) + \nabla g_i(\bar{x})' (x - \bar{x}) \leq 0 \}
\]

for \( i = 1, ..., m \), \( \nabla h_i(\bar{x})' (x - \bar{x}) = 0 \) for \( i = 1, ..., \ell \}. \] (4.17)

Now consider Examples 4.3.3, 4.3.4, and 4.3.5. In Example 4.3.3 the reader can verify that \( u_1 = u_2 = u_3 = 0 \) and \( v_1 = -8/5 \) satisfy the KKT conditions at \( \bar{x} = (4/5, 8/5)' \). In Example 4.3.4, the values of the multipliers satisfying the KKT conditions at \( \bar{x} = (2, 1)' \) are

\[ u_1 = 1/3, \quad u_2 = u_3 = 0, \quad v_1 = 2/3. \]

Finally, Example 4.3.5 does not satisfy the constraint qualification of Theorem 4.3.7 at \( \bar{x} = (1, 0)' \), since \( \nabla h_1(\bar{x}) \) and \( \nabla h_2(\bar{x}) \) are linearly dependent. In fact, no constraint qualification (known or unknown!) can hold true at this point \( \bar{x} \) because it is not a KKT point. The feasible region for the first-order linear
programming approximation LP(\(\overline{x}\)) is given by the entire \(x_1\)-axis; and unless \(\nabla f(\overline{x})\) is orthogonal to this axis, \(\overline{x}\) is not an optimal solution for LP(\(\overline{x}\)).

Theorem 4.3.8 shows that under rather mild convexity assumptions on \(f\), \(g_i\), and \(h_i\), the KKT conditions are also sufficient for local optimality. Again, we fashion this result following Theorem 4.2.16 and ask the reader to investigate other variations in Exercises 4.22 and 4.50.

4.3.8 Theorem (Karush–Kuhn–Tucker Sufficient Conditions)

Let \(X\) be a nonempty open set in \(\mathbb{R}^n\), and let \(f: \mathbb{R}^n \to \mathbb{R}\), \(g_i: \mathbb{R}^n \to \mathbb{R}\) for \(i = 1, \ldots, m\), and \(h_i: \mathbb{R}^n \to \mathbb{R}\) for \(i = 1, \ldots, \ell\). Consider Problem P:

Minimize \(f(x)\)
subject to \(g_i(x) \leq 0\) for \(i = 1, \ldots, m\)
\(h_i(x) = 0\) for \(i = 1, \ldots, \ell\)
\(x \in X\).

Let \(\overline{x}\) be a feasible solution, and let \(I = \{i : g_i(\overline{x}) = 0\}\). Suppose that the KKT conditions hold true at \(\overline{x}\); that is, there exist scalars \(\overline{u}_i \geq 0\) for \(i \in I\) and \(\overline{v}_i\) for \(i = 1, \ldots, \ell\) such that

\[
\nabla f(\overline{x}) + \sum_{i \in I} \overline{u}_i \nabla g_i(\overline{x}) + \sum_{i = 1}^{\ell} \overline{v}_i \nabla h_i(\overline{x}) = 0. \tag{4.18}
\]

Let \(J = \{i : \overline{v}_i > 0\}\) and \(K = \{i : \overline{v}_i < 0\}\). Further, suppose that \(f\) is pseudoconvex at \(\overline{x}\), \(g_i\) is quasiconvex at \(\overline{x}\) for \(i \in I\), \(h_i\) is quasiconvex at \(\overline{x}\) for \(i \in J\), and \(h_i\) is quasiconcave at \(\overline{x}\) for \(i \in K\). Then \(\overline{x}\) is a global optimal solution to Problem P. In particular, if the generalized convexity assumptions on the objective and constraint functions are restricted to the domain \(N_\varepsilon(\overline{x})\) for some \(\varepsilon > 0\), \(\overline{x}\) is a local minimum for \(P\).

**Proof**

Let \(x\) be any feasible solution to Problem P. [In case the generalized convexity assumptions hold true only by restricting the domain of the objective and constraint functions to \(N_\varepsilon(\overline{x})\), let \(x\) be any feasible solution to Problem P that also lies within \(N_\varepsilon(\overline{x})\).]

Then, for \(i \in I\), \(g_i(x) \leq g_i(\overline{x})\), since \(g_i(x) \leq 0\) and \(g_i(\overline{x}) = 0\). By the quasiconvexity of \(g_i\) at \(\overline{x}\) it follows that

\[
g_i(\overline{x} + \lambda(x - \overline{x})) = g_i(\lambda x + (1 - \lambda)\overline{x}) \leq \max\{g_i(x), g_i(\overline{x})\} = g_i(\overline{x})
\]
for all \( \lambda \in (0, 1) \). This implies that \( g_i \) does not increase by moving from \( \bar{x} \) along the direction \( x - \bar{x} \). Thus, by Theorem 4.1.2 we must have

\[
\nabla g_i(\bar{x})' (x - \bar{x}) \leq 0 \quad \text{for } i \in I. \tag{4.19}
\]

Similarly, since \( h_i \) is quasiconvex at \( \bar{x} \) for \( i \in J \), and \( h_i \) is quasiconcave at \( \bar{x} \) for \( i \in K \), we have

\[
\nabla h_i(\bar{x})' (x - \bar{x}) \leq 0 \quad \text{for } i \in J \tag{4.20}
\]

\[
\nabla h_i(\bar{x})' (x - \bar{x}) \geq 0 \quad \text{for } i \in K. \tag{4.21}
\]

Multiplying (4.19), (4.20), and (4.21) by \( \bar{u}_i \geq 0 \), \( \bar{v}_i > 0 \), and \( \bar{v}_i < 0 \), respectively, and adding, we get

\[
\left[ \sum_{i \in I} \bar{u}_i \nabla g_i(\bar{x}) + \sum_{i \in J \cup K} \bar{v}_i \nabla h_i(\bar{x}) \right]' (x - \bar{x}) \leq 0. \tag{4.22}
\]

Multiplying (4.18) by \( x - \bar{x} \) and noting that \( \bar{v}_i = 0 \) for \( i \in J \cup K \), (4.22) implies that

\[
\nabla f(\bar{x})' (x - \bar{x}) \geq 0.
\]

By the pseudoconvexity of \( f \) at \( \bar{x} \), we must have \( f(x) \geq f(\bar{x}) \), the proof is complete.

It is instructive to note that as evident from Theorem 4.3.8 and its proof, the equality constraints having positive Lagrangian multipliers at \( \bar{x} \) can be replaced by "less than or equal to" constraints, and those having negative Lagrangian multipliers at \( \bar{x} \) can be replaced by "greater than or equal to" constraints, whereas those having zero Lagrangian multipliers can be deleted and \( \bar{x} \) will still remain a KKT solution for this relaxed problem \( P' \), say. Hence, noting Theorem 4.2.16, the generalized convexity assumptions of Theorem 4.3.8 imply that \( \bar{x} \) is optimal to the relaxed problem \( P' \), and being feasible to \( P \), it is optimal for \( P \) (globally or locally as the case might be). This argument provides an alternative simpler proof for Theorem 4.3.8 based on Theorem 4.2.16. Moreover, it asserts that under generalized convexity assumptions, the sign of the Lagrangian multipliers can be used to assess whether an equality constraint is effectively behaving as a "less than or equal to" or a "greater than or equal to" constraint.

Two points of caution are worth noting here in connection with the foregoing relaxation \( P' \) of \( P \). First, under the (generalized) convexity assumptions, deleting an equality constraint that has a zero Lagrangian multiplier can create alternative optimal solutions that are not feasible to the original problem. For example, in the problem to minimize \( \{x_1 : x_1 \geq 0 \text{ and } x_2 = 1\} \), the Lagrangian
multiplier associated with the equality at the unique optimum $\mathbf{x} = (0,1)'$ is zero. However, deleting this constraint produces an infinite number of alternative optimal solutions.

Second, for the nonconvex case, note that even if $\mathbf{x}$ is optimal for $P$, it may not even be a local optimum for $P'$, although it remains a KKT point for $P'$. For example, consider the problem to minimize $(-x_1^2 - x_2^2)$ subject to $x_1 = 0$ and $x_2 = 0$. The unique optimum is obviously $\mathbf{x} = (0,0)'$, and the Lagrangian multipliers associated with the constraints at $\mathbf{x}$ are both zeros. However, deleting either of the two constraints, or even replacing either with a “less than or equal to” or “greater than or equal to” inequality, will make the problem unbounded. In general, the reader should bear in mind that deleting even nonbinding constraints for nonconvex problems can change the optimality status of a solution. Figure 4.13 illustrates one such situation. Here $g_2(x) \leq 0$ is nonbinding at the optimum $\mathbf{x}$; but deleting it changes the global optimum to the point $\hat{x}$, leaving $\mathbf{x}$ as only a local minimum. (See Exercise 4.24 for an instance in which the optimum does not even remain locally optimum after deleting a nonbinding constraint.)

**Alternative Forms of the Karush–Kuhn–Tucker Conditions for General Problems**

Consider the problem to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$, $h_i(x) = 0$ for $i = 1, \ldots, \ell$, and $x \in X$, where $X$ is an open set in $\mathbb{R}^n$. We have derived above the following necessary conditions of optimality at a feasible point $\mathbf{x}$ (under a suitable constraint qualification):

---

Figure 4.13 Caution on deleting nonbinding constraints for nonconvex problems.
\[ \nabla f(\bar{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{x}) = 0 \]

\[ u_i g_i(\bar{x}) = 0 \quad \text{for } i = 1, \ldots, m \]
\[ u_i \geq 0 \quad \text{for } i = 1, \ldots, m. \]

Some authors prefer to use the multipliers \( \lambda_i = -u_i \leq 0 \) and \( \mu_i = -v_i \). In this case, the KKT conditions can be written as follows:

\[ \nabla f(\bar{x}) - \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) - \sum_{i=1}^{\ell} u_i \nabla h_i(\bar{x}) = 0 \]

\[ \lambda_i g_i(\bar{x}) = 0 \quad \text{for } i = 1, \ldots, m \]
\[ \lambda_i \leq 0 \quad \text{for } i = 1, \ldots, m. \]

Now, consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \), \( g_i(x) \geq 0 \) for \( i = m_1 + 1, \ldots, m \), \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \), and \( x \in X \), where \( X \) is an open set in \( R^n \). Writing \( g_i(x) \geq 0 \) for \( i = m_1 + 1, \ldots, m \) as \( -g_i(x) \leq 0 \) for \( i = m_1 + 1, \ldots, m \), and using the results of Theorem 4.3.7, the necessary conditions for this problem can be expressed as follows:

\[ \nabla f(\bar{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{x}) = 0 \]

\[ u_i g_i(\bar{x}) = 0 \quad \text{for } i = 1, \ldots, m \]
\[ u_i \geq 0 \quad \text{for } i = 1, \ldots, m_1 \]
\[ u_i \leq 0 \quad \text{for } i = m_1 + 1, \ldots, m. \]

We now consider problems of the type to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \), \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \), and \( x \geq 0 \). Such problems with nonnegativity restrictions on the variables frequently arise in practice. Clearly, the KKT conditions discussed earlier would apply as usual. However, it is sometimes convenient to eliminate the Lagrangian multipliers associated with \( x \geq 0 \). The conditions then reduce to

\[ \nabla f(\bar{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{x}) \geq 0 \]

\[ \left[ \nabla f(\bar{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{x}) \right]^t \cdot x = 0 \]

\[ u_i g_i(\bar{x}) = 0 \quad \text{for } i = 1, \ldots, m \]
\[ u_i \geq 0 \quad \text{for } i = 1, \ldots, m. \]
Finally, consider the problem to maximize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$, $g_i(x) \geq 0$ for $i = m_1 + 1, \ldots, m$, $h_i(x) = 0$ for $i = 1, \ldots, \ell$, and $x \in X$, where $X$ is an open set in $\mathbb{R}^n$. The necessary conditions for optimality can be written as follows:

$$
\nabla f(\bar{x}) + \sum_{i=1}^{m_1} u_i \nabla g_i(\bar{x}) + \sum_{i=m_1+1}^{\ell} v_i \nabla h_i(\bar{x}) = 0
$$

for $i = 1, \ldots, m$

$u_i g_i(\bar{x}) = 0$

for $i = 1, \ldots, m_1$

$u_i \leq 0$

for $i = 1, \ldots, m_1$

$u_i \geq 0$

for $i = m_1 + 1, \ldots, m$.

### 4.4 Second-Order Necessary and Sufficient Conditions for Constrained Problems

In Section 4.1 we considered the unconstrained problem of minimizing $f(x)$ subject to $x \in \mathbb{R}^n$, and assuming differentiability, we derived the first-order necessary optimality condition that $\nabla f(\bar{x}) = 0$ at all local optimal solutions $\bar{x}$. However, when $\nabla f(\bar{x}) = 0$, $\bar{x}$ can be a local minimum, a local maximum, or an inflection point. To further reduce the candidate set of solutions produced by this first-order necessary optimality condition, and to assess the local optimality status of a given candidate solution, we developed second-order (and higher) necessary and/or sufficient optimality conditions.

Over Sections 4.2 and 4.3 we have developed first-order necessary optimality conditions for constrained problems. In particular, assuming a suitable constraint qualification, we have derived the first-order necessary KKT optimality conditions. Based on various (generalized) convexity assumptions, we have provided sufficient conditions to guarantee that a given solution that satisfies the first-order optimality conditions is globally or locally optimum. Analogous to the unconstrained case, we now derive second-order necessary and sufficient optimality conditions for constrained problems.

Toward this end, let us introduce the concept of a **Lagrangian function**. Consider the problem:

$$
P: \text{Minimize} \{f(x) : x \in S\}, \quad (4.23a)
$$

where

$$
S = \{x : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m, \ h_i(x) = 0 \text{ for } i = 1, \ldots, \ell, \text{ and } x \in X\}. \quad (4.23b)
$$

Assume that $f$, $g_i$ for $i = 1, \ldots, m$, and $h_i$ for $i = 1, \ldots, \ell$ are all defined on $\mathbb{R}^n \to \mathbb{R}$ and are twice differentiable, and that $X$ is a nonempty open set in $\mathbb{R}^n$. The **Lagrangian function** for this problem is defined as
As we shall learn in Chapter 6, this function enables us to formulate a duality theory for nonlinear programming problems, akin to that for linear programming problems as expounded in Theorem 2.7.3 and its corollaries. Now, let $\bar{x}$ be a KKT point for Problem P, with associated Lagrangian multipliers $\bar{u}$ and $\bar{v}$ corresponding to the inequality and equality constraints, respectively. Conditioned on $\bar{u}$ and $\bar{v}$, define the restricted Lagrangian function

$$L(x) \equiv \phi(x, \bar{u}, \bar{v}) = f(x) + \sum_{i \in I} \bar{u}_i g_i(x) + \sum_{i=1}^{\ell} \bar{v}_i h_i(x),$$

where $I = \{i: g_i(\bar{x}) = 0\}$ is the index set of the binding inequality constraints at $\bar{x}$.

Observe that the dual feasibility condition

$$\nabla f(\bar{x}) + \sum_{i \in I} \bar{u}_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} \bar{v}_i \nabla h_i(\bar{x}) = 0$$

in the KKT system is equivalent to the statement that the gradient, $\nabla L(\bar{x})$, of $L$ at $x = \bar{x}$ vanishes. Moreover, we have

$$L(x) \leq f(x) \quad \text{for all } x \in S, \quad \text{while } L(\bar{x}) = f(\bar{x})$$

because $h_i(x) = 0$ for $i = 1, \ldots, \ell$ and $g_i(x) \leq 0$ for $i \in I$ for all $x \in S$, while $\bar{u}_i g_i(\bar{x}) = 0$ for $i \in I$, and $h_i(\bar{x}) = 0$ for $i = 1, \ldots, \ell$. Hence, if $\bar{x}$ turns out to be a (local) minimizer for $L$, it will also be a (local) minimizer for Problem P. This is formalized below.

### 4.4.1 Lemma

Consider Problem P as defined in (4.23), where the objective and constraint defining functions are all twice differentiable, and where $X$ is a nonempty, open set in $R^n$. Suppose that $\bar{x}$ is a KKT point for Problem P with Lagrangian multipliers $\bar{u}$ and $\bar{v}$ associated with the inequality and equality constraints, respectively. Define the restricted Lagrangian function $L$ as in (4.25), and denote its Hessian by $\nabla^2 L$.

a. If $\nabla^2 L$ is positive semidefinite for all $x \in S$, $\bar{x}$ is a global minimum for Problem P. On the other hand, if $\nabla^2 L$ is positive semidefinite for all $x \in S \cap N_\varepsilon(\bar{x})$ for some $\varepsilon$-neighborhood $N_\varepsilon(\bar{x})$ about $\bar{x}$, $\varepsilon > 0$, $\bar{x}$ is a local minimum for Problem P.

b. If $\nabla^2 L(\bar{x})$ is positive definite, $\bar{x}$ is a strict local minimum for Problem P.
Proof

From (4.25) and (4.26), we have that \( \nabla L(\bar{x}) = 0 \). Hence, under the first condition of Part a, we obtain, by the convexity of \( L(x) \) over \( S \), that \( L(\bar{x}) \leq L(x) \) for all \( x \in S \); and thus from (4.27) we get \( f(\bar{x}) = L(\bar{x}) \leq L(x) \leq f(x) \) for all \( x \in S \). Therefore, \( \bar{x} \) solves Problem P. By restricting attention to \( S \cap N_\varepsilon(\bar{x}) \) in the second case of Part a, we conclude similarly that \( f(\bar{x}) \leq f(x) \) for all \( x \in S \cap N_\varepsilon(\bar{x}) \). This proves Part a.

Similarly, if \( \nabla^2 L(\bar{x}) \) is positive definite, by Theorem 4.1.4 since \( \nabla L(\bar{x}) = 0 \), we have that \( \bar{x} \) is a strict local minimum for \( L \). Hence, from (4.27) we deduce that \( f(\bar{x}) = L(\bar{x}) < L(x) \leq f(x) \) for all \( x \neq \bar{x} \) in \( S \cap N_\varepsilon(\bar{x}) \) for some \( \varepsilon \)-neighborhood \( N_\varepsilon(\bar{x}) \) of \( \bar{x} \), \( \varepsilon > 0 \), and this completes the proof.

The above result is related to the saddle point optimality conditions explored more fully in Chapter 6, which establish that a KKT solution \((\bar{x}, \bar{u}, \bar{v})\) for which \( \bar{x} \) minimizes \( L(x) \) subject to \( x \in S \) corresponds to a certain pair of primal and dual problems having no duality gap. Indeed, observe that the global (or local) optimality claims in Lemma 4.4.1 continue to hold true under the less restrictive assumption that \( \bar{x} \) globally (or locally) minimizes \( L(x) \) over \( S \). However, our choice of stating Lemma 4.4.1 as above is motivated by the following result, which asserts that \( d^T \nabla^2 L(\bar{x}) d \) needs to be positive only for \( d \) restricted to lie in a specified cone rather than for all \( d \in \mathbb{R}^n \) as in Lemma 4.4.1b, for us to be able to claim that \( \bar{x} \) is a strict local minimum for P. In other words, this result is shown to hold true whenever the Lagrangian function \( L(x) \) displays a positive curvature at \( \bar{x} \) along directions restricted to the set given below.

4.4.2 Theorem (KKT Second-Order Sufficient Conditions)

Consider Problem P as defined in (4.23), where the objective and constraint defining functions are all twice differentiable, and where \( X \) is a nonempty, open set in \( \mathbb{R}^n \). Let \( \bar{x} \) be a KKT point for Problem P, with Lagrangian multipliers \( \bar{u} \) and \( \bar{v} \) associated with the inequality and equality constraints, respectively. Let \( I = \{i : g_i(\bar{x}) = 0\} \), and denote \( I^+ = \{i \in I : \bar{u}_i > 0\} \) and \( I^0 = \{i \in I : \bar{u}_i = 0\} \). (\( I^+ \) and \( I^0 \) are sometimes referred to as the set of strongly active and weakly active constraints, respectively.) Define the restricted Lagrangian function \( L(x) \) as in (4.25), and denote its Hessian at \( \bar{x} \) by

\[
\nabla^2 L(\bar{x}) = \nabla^2 f(\bar{x}) + \sum_{i \in I^+} \bar{u}_i \nabla^2 g_i(\bar{x}) + \sum_{i = 1}^{\ell} \bar{v}_i \nabla^2 h_i(\bar{x}),
\]
where $\nabla^2 f(\bar{x})$, $\nabla^2 g_i(\bar{x})$ for $i \in I$, and $\nabla^2 h_i(\bar{x})$ for $i = 1, \ldots, \ell$, are the Hessians of $f$, $g_i$ for $i \in I$, and $h_i$ for $i = 1, \ldots, \ell$, respectively, all evaluated at $\bar{x}$. Define the cone
\[
C = \{d \neq 0 : \nabla g_i(\bar{x})^T d = 0 \text{ for } i \in I^+, \nabla g_i(\bar{x})^T d \leq 0 \text{ for } i \in I^0, \\
\nabla h_i(\bar{x})^T d = 0 \text{ for } i = 1, \ldots, \ell\}.
\]

Then if $d^T \nabla^2 L(\bar{x}) d > 0$ for all $d \in C$, we have that $\bar{x}$ is a strict local minimum for $P$.

**Proof**

Suppose that $\bar{x}$ is not a strict local minimum. Then, as in Theorem 4.1.4, there exists a sequence $\{x_k\}$ in $S$ converging to $\bar{x}$ such that $x_k \neq \bar{x}$ and $f(x_k) - f(\bar{x}) \leq 0$ for all $k$. Defining $d_k = (x_k - \bar{x})/\|x_k - \bar{x}\|$ and $\lambda_k = \|x_k - \bar{x}\|$ for all $k$, we have that $x_k = \bar{x} + \lambda_k d_k$, where $\|d_k\| = 1$ for all $k$, and we have that $\{\lambda_k\} \to 0^+$ as $k \to \infty$. Since $\|d_k\| = 1$ for all $k$, a convergent subsequence exists. Assume, without loss of generality, that the given sequence itself represents this convergent subsequence. Hence, $\{d_k\} \to d$, where $\|d_k\| = 1$. Moreover, we have

\[
0 \geq f(\bar{x} + \lambda_k d_k) - f(\bar{x}) = \lambda_k \nabla f(\bar{x})^T d_k + (1/2)\lambda_k^2 \nabla^2 f(\bar{x}) d_k + \lambda_k^2 \alpha_f(\bar{x}; \lambda_k d_k),
\]

(4.28a)

\[
0 \geq g_i(\bar{x} + \lambda_k d_k) - g_i(\bar{x}) = \lambda_k \nabla g_i(\bar{x})^T d_k + (1/2)\lambda_k^2 \nabla^2 g_i(\bar{x}) d_k + \lambda_k^2 \alpha_{g_i}(\bar{x}; \lambda_k d_k) \text{ for } i \in I,
\]

(4.28b)

\[
0 = h_i(\bar{x} + \lambda_k d_k) - h_i(\bar{x}) = \lambda_k \nabla h_i(\bar{x})^T d_k + (1/2)\lambda_k^2 \nabla^2 h_i(\bar{x}) d_k + \lambda_k^2 \alpha_{h_i}(\bar{x}; \lambda_k d_k) \text{ for } i = 1, \ldots, \ell,
\]

(4.28c)

where $\alpha_f$, $\alpha_{g_i}$ for $i \in I$, and $\alpha_{h_i}$ for $i = 1, \ldots, \ell$, all approach zero as $k \to \infty$.

Dividing each expression in (4.28) by $\lambda_k > 0$ and taking limits as $k \to \infty$, we obtain

\[
\nabla f(\bar{x})^T d \leq 0, \quad \nabla g_i(\bar{x})^T d \leq 0 \text{ for } i \in I \quad \text{and} \\
\nabla h_i(\bar{x})^T d = 0 \text{ for } i = 1, \ldots, \ell.
\]

(4.29)

Now, since $\bar{x}$ is a KKT point, we have $\nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{x}) = 0$. Taking the inner product of this with $d$ and using (4.29), we conclude that
The Fritz John and Karush-Kuhn-Tucker Optimality Conditions

\[ \nabla f(\bar{x})'d = 0, \quad \nabla g_i(\bar{x})'d = 0 \text{ for } i \in I^+, \quad \nabla g_i(\bar{x})'d \leq 0 \text{ for } i \in I^0, \]

and

\[ \nabla h_i(\bar{x})'d = 0 \text{ for } i = 1, \ldots, \ell. \quad (4.30) \]

Hence, in particular, \( d \in C \). Furthermore, multiplying each of (4.28b) by \( \bar{u}_i \) for \( i \in I \), and each of (4.28c) by \( \bar{v}_i \) for \( i = 1, \ldots, \ell \), and adding, we get, using \( \nabla f(\bar{x})'d_k + \sum_{i \in I} \bar{u}_i \nabla g_i(\bar{x})'d_k + \sum_{i=1}^{\ell} \bar{v}_i \nabla h_i(\bar{x})'d_k = 0, \]

\[ 0 \geq \frac{\lambda_k^2}{2} d_k' \nabla^2 L(\bar{x})d_k + \lambda_k^2 \left[ \alpha_f(\bar{x}; \lambda_k d_k) + \sum_{i \in I} \bar{u}_i \alpha_{g_i}(\bar{x}; \lambda_k d_k) + \sum_{i=1}^{\ell} \bar{v}_i \alpha_{h_i}(\bar{x}; \lambda_k d_k) \right]. \]

Dividing the above inequality by \( \lambda_k^2 > 0 \) and taking limits as \( k \to \infty \), we obtain \( d' \nabla^2 L(\bar{x})d \leq 0 \), where \( \|d\| = 1 \) and \( d \in C \). This is a contradiction. Therefore, \( \bar{x} \) must be a strict local minimum for Problem P, and the proof is complete.

**Corollary**

Consider Problem P as defined in the theorem, and let \( \bar{x} \) be a KKT point with associated Lagrangian multipliers \( \bar{u} \) and \( \bar{v} \) corresponding to the inequality and equality constraints, respectively. Furthermore, suppose that the collection of vectors \( \nabla g_i(\bar{x}) \) for \( i \in I^+ = \{i \in I : \bar{u}_i > 0\} \) and \( \nabla h_i(\bar{x}) \) for \( i = 1, \ldots, \ell \) contains a set of \( n \) linearly independent vectors. Then \( \bar{x} \) is a strict local minimum for P.

**Proof**

Under the stated linear independence condition of the corollary, we have that \( C = \emptyset \), so Theorem 4.4.2 holds true vacuously by default. This completes the proof.

Several remarks concerning Theorem 4.4.2 are in order at this point. First, observe that it might appear from the proof of the theorem that the result can be strengthened by further restricting the cone \( C \) to include the constraint \( \nabla f(\bar{x})'d = 0 \). Although this is valid, it does not further restrict \( C \), since when \( \bar{x} \) is a KKT point and \( d \in C \), we have automatically that \( \nabla f(\bar{x})'d = 0 \). Second, observe that if the problem is unconstrained, Theorem 4.2.2 reduces to asserting that if \( \nabla f(\bar{x}) = 0 \) and if \( \nabla^2 f(\bar{x}) \equiv H(\bar{x}) \) is positive definite, \( \bar{x} \) is a strict local minimum. Hence, Theorem 4.1.4 is a special case of this result. Similarly, Lemma 4.4.1b is a special case of this result. Finally, observe that for linear programming problems, this sufficient condition does not necessarily hold true,
except under the condition of the corollary, whence $\bar{x}$ is a unique extreme point optimal solution.

We now turn our attention to the counterpart of Theorem 4.4.2 that deals with second-order necessary optimality conditions. Theorem 4.4.3 shows that if $\bar{x}$ is a local minimum, then under a suitable second-order constraint qualification, it is a KKT point; and moreover, $d^T\nabla^2 L(\bar{x})d \geq 0$ for all $d$ belonging to $C$ as defined in Theorem 4.4.2. The last statement indicates that the Lagrangian function $L$ has a nonnegative curvature at $\bar{x}$ along any direction in $C$.

4.4.3 Theorem (KKT Second-Order Necessary Conditions)

Consider Problem P as defined in (4.23), where the objective and constraint defining functions are all twice differentiable, and where $X$ is a nonempty, open set in $\mathbb{R}^n$. Let $\bar{x}$ be a local minimum for Problem P, and denote $I = \{i: g_i(\bar{x}) = 0\}$. Define the restricted Lagrangian function $L(x)$ as in (4.25), and denote its Hessian at $\bar{x}$ by

$$
\nabla^2 L(\bar{x}) = \nabla^2 f(\bar{x}) + \sum_{i \in I} \bar{u}_i \nabla^2 g_i(\bar{x}) + \sum_{i=1}^\ell \bar{v}_i \nabla^2 h_i(\bar{x}),
$$

where $\nabla^2 f(\bar{x})$, $\nabla^2 g_i(\bar{x})$ for $i \in I$, and $\nabla^2 h_i(\bar{x})$ for $i = 1, \ldots, \ell$, are the Hessians of $f$, $g_i$ for $i \in I$, and $h_i$ for $i = 1, \ldots, \ell$, respectively, all evaluated at $\bar{x}$. Assume that $\nabla g_i(\bar{x})$ for $i \in I$, and $\nabla h_i(\bar{x})$ for $i = 1, \ldots, \ell$, are linearly independent. Then $\bar{x}$ is a KKT point having Lagrange multipliers $\bar{u}_i \geq 0$ and $\bar{v}_i$ associated with the inequality and the equality constraints, respectively. Moreover, $d^T\nabla^2 L(\bar{x})d \geq 0$ for all $d \in C = \{d \neq 0: \nabla g_i(\bar{x})'d = 0 \text{ for } i \in I_1, \nabla g_i(\bar{x})'d \leq 0 \text{ for } i \in I_0, \nabla h_i(\bar{x})'d = 0 \text{ for all } i = 1, \ldots, \ell\}$, where $I_1 = \{i \in I: \bar{u}_i > 0\}$ and $I_0 = \{i \in I: \bar{u}_i = 0\}$.

**Proof**

By Theorem 4.3.7 we have directly that $\bar{x}$ is a KKT point. Now, if $C = \emptyset$, the result is trivially true. Otherwise, consider any $d \in C$, and denote $I(d) = \{i \in I: \nabla g_i(\bar{x})'d = 0\}$. For $\lambda \geq 0$, define $\alpha: \mathbb{R} \to \mathbb{R}^n$ by the following differential equation and boundary condition:

$$
\frac{d\alpha(\lambda)}{d\lambda} = P(\lambda)d, \quad \alpha(0) = \bar{x},
$$

where $P(\lambda)$ is the matrix that projects any vector in the null space of the matrix having rows $\nabla g_i(\alpha(\lambda))$, $i \in I(d)$, and $\nabla h_i(\alpha(\lambda))$, $i = 1, \ldots, \ell$. Following the proof of Theorem 4.3.1 [by treating $g_i$ for $i \in I(d)$, and $h_i$, $i = 1, \ldots, \ell$, as the "equations" therein, and treating $g_i$ for $i \in I - I(d)$, for which $\nabla g_i(\bar{x})'d < 0$, as
the "inequalities" therein], we obtain that \( \alpha(\lambda) \) is feasible for \( 0 \leq \lambda \leq \delta \), for some \( \delta > 0 \).

Now, consider a sequence \( \{ \lambda_k \} \to 0^+ \) and denote \( x_k = \alpha(\lambda_k) \) for all \( k \). By the Taylor series expansion, we have

\[
L(x_k) = L(\bar{x}) + \nabla L(\bar{x})^t(x_k - \bar{x}) + \frac{1}{2}(x_k - \bar{x})^t \nabla^2 L(\bar{x})(x_k - \bar{x})
+ \left\| x_k - \bar{x} \right\|^2 \beta[\bar{x};(x_k - \bar{x})],
\]

where \( \beta[\bar{x};(x_k - \bar{x})] \to 0 \) as \( x_k \to \bar{x} \). Since \( g_i(x_k) = 0 \) for all \( i \in I(d) \supseteq I^+ \) and \( h_i(x_k) = 0 \) for all \( i = 1, \ldots, \ell \), we have that \( L(x_k) = f(x_k) \) from (4.25). Similarly, \( L(\bar{x}) = f(\bar{x}) \). Also, since \( \bar{x} \) is a KKT point, we have that \( \nabla L(\bar{x}) = 0 \). Moreover, since \( x_k = \alpha(\lambda_k) \) is feasible, \( x_k \to \bar{x} \) as \( \lambda_k \to 0^+ \) or as \( k \to \infty \), and since \( \bar{x} \) is a local minimum, we must have \( f(x_k) \geq f(\bar{x}) \) for \( k \) sufficiently large. Consequently, from (4.31), we get

\[
\frac{f(x_k) - f(\bar{x})}{\lambda_k^2} = \frac{1}{2} \frac{(x_k - \bar{x})^t}{\lambda_k} \nabla^2 L(\bar{x}) \frac{x_k - \bar{x}}{\lambda_k}
+ \left\| \frac{x_k - \bar{x}}{\lambda_k} \right\|^2 \beta[\bar{x};(x_k - \bar{x})] \geq 0
\]

for \( k \) large enough. But note that

\[
\lim_{k \to \infty} \frac{x_k - \bar{x}}{\lambda_k} = \lim_{k \to \infty} \frac{\alpha(\lambda_k) - \alpha(0)}{\lambda_k} = \alpha'(0) = P(0)d = d
\]

since \( d \) is already in the null space of the matrix having rows \( \nabla g_i(\bar{x}) \) for \( i \in I(d) \), and \( \nabla h_i(\bar{x}) \) for \( i = 1, \ldots, \ell \). Taking limits in (4.32a) as \( k \to \infty \) and using (4.32b), we get that \( d^t \nabla^2 L(\bar{x})d \geq 0 \), and this completes the proof.

Observe that the set \( C \) defined in the theorem is a subset of \( G_0 \cap H_0 \) and that the nonnegative curvature of \( L \) at \( \bar{x} \) is required for all \( d \in C \), but not necessarily for all \( d \in G_0 \cap H_0 \). Furthermore, note that if the problem is unconstrained, Theorem 4.4.3 reduces to asserting that \( \nabla f(\bar{x}) = 0 \) and \( H(\bar{x}) \) is positive semidefinite at a local minimum \( \bar{x} \). Hence, Theorem 4.1.3 is a special case of this result. Let us now illustrate the use of the foregoing results.

**4.4.4 Example (McCormick [1967])**

Consider the nonconvex programming problem
P: Minimize\{\((x_1 - 1)^2 + x_2^2 : g_1(x) = 2kx_1 - x_2^2 \leq 0\}\),

where \(k\) is a positive constant. Figure 4.14 illustrates two possible ways in which the optimum is determined, depending on the value of \(k\).

Note that \(\nabla g_1(x) = (2k, -2x_2)' \neq (0, 0)'\), and hence the linear independence constraint qualification holds true at any feasible solution \(x\). The KKT conditions require primal feasibility and that

\[
\begin{bmatrix}
2(x_1 - 1) \\
2x_2
\end{bmatrix} + u_1 \begin{bmatrix}
2k \\
-2x_2
\end{bmatrix} = \begin{bmatrix}
0 \\
0
\end{bmatrix},
\]

where \(u_1 \geq 0\) and \(u_1[2kx_1 - x_2^2] = 0\). If \(u_1 = 0\), we must have \((x_1, x_2) = (1, 0)\), which is the unconstrained minimum and which is infeasible for any \(k > 0\). Hence, \(u_1\) must be positive for any KKT point; so, by complementary slackness, \(2kx_1 = x_2^2\) must hold true. Furthermore, by the second dual feasibility constraint, we must either have \(x_2 = 0\) or \(u_1 = 1\). If \(x_2 = 0\), \(2kx_1 = x_2^2\) yields \(x_1 = 0\), and the first dual feasibility constraint yields \(u_1 = 1/k\). This gives one KKT solution. Similarly, from the KKT conditions, when \(u_1 = 1\), we obtain that \(x_1 = 1 - k\) and \(x_2 = \pm\sqrt{2k(1-k)}\), which yields a different set of KKT solutions when \(0 < k < 1\). Hence, the KKT solutions are \(\{\bar{x}_1 = (0,0)', \bar{u}_1 = 1/k\}\) for any \(k > 0\), and

![Figure 4.14 Two cases of optimal solutions: Example 4.4.4.](image-url)
The Fritz John and Karush–Kuhn–Tucker Optimality Conditions

\{\bar{x}^2 = (1 - k, \sqrt{2k(1 - k)}, \bar{u}_1^2 = 1)\} along with \{\bar{x}^3 = (1 - k, -\sqrt{2k(1 - k)}, \bar{u}_1^3 = 1)\} whenever \(0 < k < 1\).

By examining convex combinations of objective values at the above KKT points and at any other point \(x\) on the constraint surface, for example, it is readily verified that \(g_I\) is not quasiconvex at these points; and thus, while the first-order necessary condition of Theorem 4.2.13 is satisfied, the sufficient condition of Theorem 4.2.16 does not hold true. Hence, we are uncertain about the character of the above KKT solutions using these results.

Now, let us examine the second-order necessary condition of Theorem 4.4.3. Note that \(L(x) = f(x) + \bar{u}g(x) = (x_1 - 1)^2 + x_2^2 + \bar{u}[2kx_1 - x_2^2]\), so

\[\nabla^2 L(\bar{x}) = \begin{bmatrix} 2 & 0 \\ 0 & 2(1 - \bar{u}) \end{bmatrix}.\]

Furthermore, the cone \(C\) defined in Theorem 4.4.3 is given by (since \(\bar{u}_1 > 0\) at any KKT point)

\[C = \{d \neq 0 : kd_1 = \bar{x}_2 d_2\}.\]

For the KKT solution \((\bar{x}^1, \bar{u}_1^1)\), Theorem 4.4.3 requires that \(2d_1^2 + 2(1 - 1/k)d_2^2 \geq 0\) for all \((d_1, d_2)\) such that \(d_1 = 0\). Whenever \(k \geq 1\), this obviously holds true. However, when \(0 < k < 1\), this condition is violated. Hence, using Theorem 4.4.3, we can conclude that \(\bar{x}^1\) is not a local minimum for \(0 < k < 1\). On the other hand, since \(\bar{u}_1^2 = \bar{u}_1^3 = 1\), \(\nabla^2 L(\bar{x}^2)\) and \(\nabla^2 L(\bar{x}^3)\) are positive semidefinite, and hence the other sets of KKT solutions satisfy the second-order necessary optimality conditions.

Next, let us examine the second-order sufficient conditions of Theorem 4.4.2. For the KKT solution \((\bar{x}^1, \bar{u}_1^1)\), whenever \(k > 1\), \(\nabla^2 L(\bar{x}^1)\) is itself positive definite; so, even by Lemma 4.4.1b, we have that \(\bar{x}^1\) is a strict local minimum. However, for \(k = 1\), although \(\bar{x}^1\) solves Problem \(P\), we are unable to recognize this via Theorem 4.4.2, since \(d^t \nabla^2 L(\bar{x}^1)d = 2d_1^2 = 0\) for \(d \in C = \{d \neq 0 : d_1 = 0\}\).

Next, consider the KKT solution \((\bar{x}^2, \bar{u}_1^2)\) for \(0 < k < 1\). Here \(C = \{d \neq 0 : kd_1 = \sqrt{2k(1 - k)d_2}\}\); and for any \(d \in C\), we have \(d^t \nabla^2 L(\bar{x}^2)d = 2d_1^2 > 0\). Hence, by Theorem 4.4.2, \(\bar{x}^2\) is a strict local minimum for \(0 < k < 1\). Note that \(\nabla^2 L(\bar{x}^2)\) itself is not positive definite, and therefore Theorem 4.4.2 plays a critical role in concluding the local minimum status of \(\bar{x}^2\). Similarly, \(\bar{x}^3\) is a strict local minimum for \(0 < k < 1\). The global minimum status of the strict local minima above must be established by other means because of the nonconvexity of the problem (see Exercise 4.40).
Exercises

[4.1] Consider the univariate function \( f(x) = xe^{-2x} \). Find all local minima/maxima and inflection points. Also, what can you claim about a global minimum and a global maximum for \( f \)? Give analytical justifications for your claims.

[4.2] Consider the following linear program:

Maximize \( x_1 + 3x_2 \)
subject to \( 2x_1 + 3x_2 \leq 6 \)
\( -x_1 + 4x_2 \leq 4 \)
\( x_1, x_2 \geq 0. \)

a. Write the KKT optimality conditions.

b. For each extreme point, verify whether or not the KKT conditions hold true, both algebraically and geometrically. From this, find an optimal solution.

[4.3] Consider the following problem:

Minimize \( x_1^2 + 2x_2^2 \)
subject to \( x_1 + x_2 - 2 = 0. \)

Find a point satisfying the KKT conditions and verify that it is indeed an optimal solution. Re-solve the problem if the objective function is replaced by \( x_1^3 + x_2^3 \).

[4.4] Consider the following unconstrained problem:

Minimize \( 2x_1^2 - x_1x_2 + x_2^2 - 3x_1 + e^{2x_1 + x_2} \).

a. Write the first-order necessary optimality conditions. Is this condition also sufficient for optimality? Why?

b. Is \( \bar{x} = (0, 0)' \) an optimal solution? If not, identify a direction \( d \) along which the function would decrease.

c. Minimize the function starting from \( (0, 0) \) along the direction \( d \) obtained in Part b.

d. Dropping the last term in the objective function, use a classical direct optimization technique to solve this problem.

[4.5] Consider the following problem:

Minimize \( x_1^4 + x_2^4 + 12x_1^2 + 6x_2^2 - x_1x_2 - x_1 - x_2 \)
subject to \( x_1 + x_2 \geq 6 \)
\( 2x_1 - x_2 \geq 3 \)
\( x_1 \geq 0, x_2 \geq 0. \)
Write out the KKT conditions and show that \((x_1, x_2) = (3, 3)\) is the unique optimal solution.

[4.6] Consider the problem to minimize \(\|Ax - b\|^2\), where \(A\) is an \(m \times n\) matrix and \(b\) is an \(m\)-vector.

a. Give a geometric interpretation of the problem.
b. Write a necessary condition for optimality. Is this also a sufficient condition?
c. Is the optimal solution unique? Why or why not?
d. Can you give a closed-form solution of the optimal solution? Specify any assumptions that you may need.
e. Solve the problem for

\[
A = \begin{bmatrix}
2 & -1 & 0 \\
0 & 2 & 2 \\
0 & 1 & 0 \\
1 & 0 & 1
\end{bmatrix}, \quad b = \begin{bmatrix}
2 \\
6 \\
2 \\
0
\end{bmatrix}.
\]

[4.7] Consider the following problem:

Minimize \(\left(x_1 - \frac{9}{4}\right)^2 + (x_2 - 2)^2\)

subject to \(x_2 - x_1^2 \geq 0\)
\(x_1 + x_2 \leq 6\)
\(x_1, x_2 \geq 0.\)

a. Write the KKT optimality conditions and verify that these conditions hold true at the point \(\bar{x} = (3/2, 9/4)^t\).
b. Interpret the KKT conditions at \(\bar{x}\) graphically.
c. Show that \(\bar{x}\) is indeed the unique global optimal solution.

[4.8] Consider the following problem:

Minimize \(\frac{x_1 + 3x_2}{2x_1 + x_2} + 3\)

subject to \(2x_1 + x_2 \leq 12\)
\(-x_1 + 2x_2 \leq 4\)
\(x_1, x_2 \geq 0.\)

a. Show that the KKT conditions are sufficient for this problem.
b. Show that any point on the line segment joining the points \((0, 0)\) and \((6, 0)\) is an optimal solution.

[4.9] Consider the following problem, where \(e\) is a nonzero vector in \(\mathbb{R}^n\):
Maximize $c'd$

subject to $d'd \leq 1$.

a. Show that $\bar{d} = c/\|c\|$ is a KKT point. Furthermore, show that $\bar{d}$ is indeed the unique global optimal solution.

b. Using the result of Part a, show that the direction of steepest ascent of $f$ at a point $x$ is given by $\nabla f(x)/\|\nabla f(x)\|$ provided that $\nabla f(x) \neq 0$.

[4.10] Consider the problem to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$.

a. Show that verifying whether a point $\bar{x}$ is a KKT point is equivalent to finding a vector $u$ satisfying a system of the form $A'u = c$, $u \geq 0$. (This can be done using Phase I of linear programming.)

b. Indicate the modifications needed in Part a if the problem had equality constraints.

c. Illustrate Part a by the following problem, where $\bar{x} = (1, 2, 5)'$:

Minimize $2x_1^2 + x_2^2 + 2x_3^2 + x_1x_3 - x_1x_2 + x_1 + 2x_3$

subject to $x_1^2 + x_2^2 - x_3 \leq 0$

$x_1 + x_2 + 2x_3 \leq 16$

$x_1 + x_2 \geq 3$

$x_1, x_2, x_3 \geq 0$.

[4.11] Consider the problem to minimize $f(x_1, x_2) = (x_2^2 - x_1)(x_2^2 - 2x_1)$, and let $\bar{x} = (0, 0)'$. Show that for each $d \in \mathbb{R}^n$, $\|d\| = 1$, there exists a $\delta_d > 0$ such that for $-\delta_d \leq \lambda \leq \delta_d$, we have $f(\bar{x} + \lambda d) \geq f(\bar{x})$. However, show that $\inf\{\delta_d : d \in \mathbb{R}^n, \|d\| = 1\} = 0$. In reference to Figure 4.1, discuss what this entails regarding the local optimality of $\bar{x}$.

[4.12] Consider the following problem, where $a_j$, $b$, and $c_j$ are positive constants:

Minimize $\sum_{j=1}^{n} \frac{c_j}{x_j}$

subject to $\sum_{j=1}^{n} a_jx_j = b$

$x_j \geq 0$ for $j = 1, \ldots, n$.

Write the KKT conditions and solve for the point $\bar{x}$ satisfying these conditions.

[4.13] Consider Problem P to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$ and $h_i(x) = 0$ for $i = 1, \ldots, \ell$. Suppose that this problem is reformulated as $\bar{P}$:
Minimize \( \{ f(x): g_i(x) + s_i^2 = 0 \text{ for } i = 1, \ldots, m \text{ and } h_i(x) = 0 \text{ for } i = 1, \ldots, \ell \} \).

Write the KKT conditions for \( P \) and for \( \overline{P} \) and compare them. Explain any difference between the two and what arguments you can use to resolve them. Express your opinion on using the formulation \( \overline{P} \) to solve the problem.

**[4.14]** Consider Problem \( P \) to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \text{ for } i = 1, \ldots, m \) and \( h_i(x) = 0 \text{ for } i = 1, \ldots, \ell \). Show that \( P \) is mathematically equivalent to the following single constraint problem \( \overline{P} \), where \( s_1, \ldots, s_m \) are additional variables:

\[
\overline{P}: \text{Minimize } \left\{ f(x): \sum_{i=1}^m [g_i(x) + s_i^2]^2 + \sum_{i=1}^\ell h_i^2(x) = 0 \right\}.
\]

Write out the FJ and the KKT conditions for \( \overline{P} \). What statements can you make about the relationship between local optima, FJ, and KKT points? What is your opinion about the utility of \( \overline{P} \) in solving \( P \)?

**[4.15]** In geometric programming, the following result is used. If \( x_1, \ldots, x_n \geq 0 \),

\[
\frac{1}{n} \sum_{j=1}^n x_j \geq \left( \prod_{j=1}^n x_j \right)^{1/n}.
\]

Prove the result using the KKT conditions.

*Hint: Consider one of the following problems and justify your use of it:

- Minimize \( \sum_{j=1}^n x_j \)

subject to \( \prod_{j=1}^n x_j = 1, x_j \geq 0 \text{ for } j = 1, \ldots, n \).

- Maximize \( \prod_{j=1}^n x_j \)

subject to \( \sum_{j=1}^n x_j = 1, x_j \geq 0 \text{ for } j = 1, \ldots, n \].

**[4.16]** Consider the quadratic assignment program to minimize \( c'x + (1/2)x'Qx \) subject to the assignment constraints \( \sum_{j=1}^m x_{ij} = 1 \text{ for all } i = 1, \ldots, m \), \( \sum_{i=1}^m x_{ij} = 1 \) for all \( j = 1, \ldots, m \), and \( x \) is binary valued. Here the component \( x_{ij} \) of the vector \( x \) takes on a value of 1 if \( i \) is assigned to \( j \) and is 0 otherwise, for \( i, j = 1, \ldots, m \).

Show that if \( M \) exceeds the sum of absolute values of elements in any row of \( Q \), the matrix \( \overline{Q} \) obtained from \( Q \) by subtracting \( M \) from each diagonal element is negative definite. Now, consider the following problem:
Using the fact that the extreme points of \( \overline{\text{QAP}} \) are all binary valued, show that \( \overline{\text{QAP}} \) is equivalent to QAP. Moreover, show that every extreme point of \( \overline{\text{QAP}} \) is a KKT point. (This exercise is due to Bazaraa and Sherali [1982].)

[4.17] Answer the following and justify your answer:

a. For a minimization nonlinear program, can a KKT point be a local maximum?

b. Let \( f \) be differentiable, let \( X \) be convex, and let \( \overline{x} \in X \) satisfy
\[
\forall f'(\overline{x})(x - \overline{x}) > 0 \quad \text{for all } x \in X, \quad x \neq \overline{x}.
\]
Is \( \overline{x} \) necessarily a local minimum?

c. What is the effect on application of the FJ and KKT optimality conditions by duplicating an equality constraint or an inequality constraint in the problem?

[4.18] Write the KKT necessary optimality conditions for Exercises 1.3 and 1.4. Using these conditions, find the optimal solutions.

[4.19] Let \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) be infinitely differentiable, and let \( \overline{x} \in \mathbb{R}^n \). For any \( d \in \mathbb{R}^n \), define
\[
F_d(\lambda) = f(\overline{x} + \lambda d) \quad \text{for } \lambda \in \mathbb{R}.
\]
Write out the infinite Taylor series expansion for \( F_d(\lambda) \) and compute \( F_d'(\lambda) \). Compare the nonnegativity or positivity of this expression with the necessary and sufficient Taylor series-based inequality for \( \overline{x} \) to be a local minimum for \( f \). What conclusions can you draw?

[4.20] Consider the following one-dimensional minimization problem:

\[
\begin{align*}
\text{Minimize} \quad & f(x + \lambda d) \\
\text{subject to} \quad & \lambda \geq 0,
\end{align*}
\]

where \( x \) is a given vector and \( d \) is a given nonzero direction.

a. Write a necessary condition for a minimum if \( f \) is differentiable. Is this condition also sufficient? If not, what assumptions on \( f \) would make the necessary condition also sufficient?

b. Suppose that \( f \) is convex but not differentiable. Can you develop a necessary optimality condition for the above problem using subgradients of \( f \) as defined in Section 3.2?

[4.21] Use the KKT conditions to prove Farkas's theorem discussed in Section 2.3. (Hint: Consider the problem to maximize \( c'x \) subject to \( Ax \leq 0 \).)

[4.22] Suppose that \( f: \mathbb{R} \rightarrow \mathbb{R} \), where \( S \subseteq \mathbb{R}^n \).

a. If \( f \) is pseudoconvex over \( N_\epsilon(\overline{x}) \cap S \), does this imply that \( f \) is pseudoconvex at \( \overline{x} \)?
b. If \( f \) is strictly pseudoconvex at \( \vec{x} \), does this imply that \( f \) is quasiconvex at \( \vec{x} \)?

c. For each of the FJ and KKT sufficiency theorems for both the equality and equality–inequality constraint cases, provide alternative sets of sufficient conditions to guarantee local optimality of a point satisfying these conditions. Prove your claims. Examine your proof for possibly strengthening the theorem by weakening your assumptions.

\[ \textbf{4.23} \] Let \( X \) be a nonempty open set in \( \mathbb{R}^n \), and consider \( f: \mathbb{R}^n \to \mathbb{R} \), \( g_i: \mathbb{R}^n \to \mathbb{R} \) for \( i = 1, \ldots, m \), and \( h_i: \mathbb{R}^n \to \mathbb{R} \), for \( i = 1, \ldots, \ell \). Consider Problem P:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m \\
& \quad h_i(x) = 0 \quad \text{for } i = 1, \ldots, \ell \\
& \quad x \in X.
\end{align*}
\]

Let \( \vec{x} \) be a feasible solution, and let \( l = \{ i : g_i(\vec{x}) = 0 \} \). Suppose that the KKT conditions hold true at \( \vec{x} \); that is, there exist scalars \( \vec{u}_i \geq 0 \) for \( i \in l \) and \( \vec{v}_i \) for \( i = 1, \ldots, \ell \) such that

\[
\nabla f(\vec{x}) + \sum_{i \in l} \vec{u}_i \nabla g_i(\vec{x}) + \sum_{i=1}^{\ell} \vec{v}_i \nabla h_i(\vec{x}) = 0.
\]

a. Suppose that \( f \) is pseudoconvex at \( \vec{x} \) and that \( \phi \) is quasiconvex at \( \vec{x} \), where

\[
\phi(x) = \sum_{i \in l} \vec{u}_i g_i(x) + \sum_{i=1}^{\ell} \vec{v}_i h_i(x).
\]

Show that \( \vec{x} \) is a global optimal solution to Problem P.

b. Show that if \( f + \sum_{i \in l} \vec{u}_i g_i + \sum_{i=1}^{\ell} \vec{v}_i h_i \) is pseudoconvex, \( \vec{x} \) is a global optimal solution to Problem P.

c. Show by means of examples that the convexity assumptions in Parts a and b and those of Theorem 4.3.8 are not equivalent to each other.

d. Relate this result to Lemma 4.4.1 and to the discussion immediately following it.

\[ \textbf{4.24} \] Let \( \vec{x} \) be an optimal solution to the problem of minimizing \( f(x) \) subject to \( g_i(x) \leq 0 \), \( i = 1, \ldots, m \) and \( h_i(x) = 0 \), \( i = 1, \ldots, \ell \). Suppose that \( g_k(\vec{x}) < 0 \) for some \( k \in \{1, \ldots, m\} \). Show that if this nonbinding constraint is deleted, it is possible that \( \vec{x} \) is not even a local minimum for the resulting problem. [\textit{Hint: Consider } g_k(\vec{x}) = -1 \text{ and } g_k(x) = 1 \text{ for } x \neq \vec{x}. ] \] Show that if all problem-defining functions are continuous, then, by deleting nonbinding constraints, \( \vec{x} \) remains at least a local optimal solution.
Consider the bilinear program to minimize $c^t x + d^t y + x^t H y$ subject to $x \in X$ and $y \in Y$, where $X$ and $Y$ are bounded polyhedral sets in $R^n$ and $R^m$, respectively. Let $\hat{x}$ and $\hat{y}$ be extreme points of the sets $X$ and $Y$, respectively.

a. Verify that the objective function is neither quasiconvex nor quasiconcave.

b. Prove that there exists an extreme point $(\hat{x}, \hat{y})$ that solves the bilinear program.

c. Prove that the point $(\hat{x}, \hat{y})$ is a local minimum of the bilinear program if and only if the following are true: (i) $c^t (x - \hat{x}) \geq 0$ and $d^t (y - \hat{y}) \geq 0$ for each $x \in X$ and $y \in Y$; (ii) $c^t (x - \hat{x}) + d^t (y - \hat{y}) > 0$ whenever $(x - \hat{x})^t H (y - \hat{y}) < 0$.

d. Show that the point $(\hat{x}, \hat{y})$ is a KKT point if and only if $(c^t + \hat{y}^t H)(x - \hat{x}) \geq 0$ for each $x \in X$ and $(d^t + \hat{x}^t H)(y - \hat{y}) \geq 0$ for each $y \in Y$.

e. Consider the problem to minimize $x_2 + y_1 + x_2 y_1 - x_1 y_2 + x_2 y_2$ subject to $(x_1, x_2) \in X$ and $(y_1, y_2) \in Y$, where $X$ is the polyhedral set defined by its extreme points $(0, 0), (0, 1), (1, 4), (2, 4), (3, 0)$, and $Y$ is the polyhedral set defined by its extreme points $(0, 0), (0, 1), (1, 5), (3, 5), (4, 4), (3, 0)$. Verify that the point $(x_1, x_2, y_1, y_2) = (0, 0, 0, 0)$ is a KKT point but not a local minimum. Verify that the point $(x_1, x_2, y_1, y_2) = (3, 0, 1, 5)$ is both a KKT point and a local minimum. What is the global minimum to the problem?

Consider the problem to minimize $f(x)$ subject to $x \geq 0$, where $f$ is a differentiable convex function. Let $\bar{x}$ be a given point and denote $Vf(\bar{x})$ by $(V_1, \ldots, V_n)^t$. Show that $\bar{x}$ is an optimal solution if and only if $d = 0$, where $d$ is defined by

$$d_i = \begin{cases} -V_i & \text{if } x_i > 0 \text{ or } V_i < 0 \\ 0 & \text{if } x_i = 0 \text{ and } V_i \geq 0. \end{cases}$$

Consider the problem to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$. Let $\bar{x}$ be a feasible point, and let $I = \{i : g_i(\bar{x}) = 0\}$. Suppose that $f$ is differentiable at $\bar{x}$ and that each $g_i$ for $i \in I$ is differentiable and concave at $\bar{x}$. Furthermore, suppose that each $g_i$ for $i \not\in I$ is continuous at $\bar{x}$. Consider the following linear program:
Minimize $Vf(x)'d$
subject to $Vg_i(x)'d \leq 0$ for $i \in I$

$-1 \leq d_j \leq 1$ for $j = 1, ..., n.$

Let $\bar{d}$ be an optimal solution with objective function value $\bar{z}$.

a. Show that $\bar{z} \leq 0$.

b. Show that if $\bar{z} < 0$, there exists a $\delta > 0$ such that $\bar{x} + \lambda \bar{d}$ is feasible and $f(\bar{x} + \lambda \bar{d}) < f(\bar{x})$ for each $\lambda \in (0, \delta)$.

c. Show that if $\bar{z} = 0$, $\bar{x}$ satisfies the KKT conditions.

[4.28] Consider the following problem, where $y$, $e$, and $y_0$ belong to $R^n$, and where $y = (y_1, ..., y_n)'$, $y_0 = (1/n, ..., 1/n)'$, and $e = (1, ..., 1)'.$

Minimize $\{y : \|y - y_0\|^2 \leq 1/n(n - 1), \ e'y = 1\}$

Interpret this problem with respect to an inscribed sphere in the simplex defined by $\{y : e'y = 1, \ y \geq 0\}$. Write the KKT conditions for this problem and verify that $(0, 1/(n - 1), ..., 1/(n - 1))$ is an optimal solution.

[4.29] Let $f : R^m \rightarrow R$, $g_i : R^m \rightarrow R$ for $i = 1, ..., m$ be convex functions. Consider the problem to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, ..., m$. Let $M$ be a proper subset of $\{1, ..., m\}$, and suppose that $\hat{x}$ solves the problem to minimize $f(\hat{x})$ subject to $g_i(\hat{x}) \leq 0$ for $i \in M$. Let $V = \{i : g_i(\hat{x}) > 0\}$. If $\bar{x}$ solves the original problem, and if $f(\bar{x}) > f(\hat{x})$, show that $g_i(\bar{x}) = 0$ for some $i \in V$. Show that this is not necessarily true if $f(\bar{x}) = f(\hat{x})$. (This exercise also shows that if an unconstrained minimum of $f$ is infeasible and has an objective value less than the optimum value, any constrained minimum lies on the boundary of the feasible region.)

[4.30] Consider Problem $P$, to minimize $f(x)$ subject to some “less than or equal to” type of linear inequality constraints. Let $\bar{x}$ be a feasible solution, and let the binding constraints be represented as $Ax = b$, where $A$ is an $m \times n$ matrix of rank $m$. Let $d = -Vf(\bar{x})$ and consider the following problem.

$\bar{P}:$ Minimize $\left\{\frac{1}{2}\|x - (\bar{x} + d)\|^2 : Ax = b \right\}$.

Let $\hat{x}$ solve $\bar{P}$.

a. Provide a geometric interpretation for $\bar{P}$ and its solution $\hat{x}$.

b. Write the KKT conditions for $\bar{P}$. Discuss whether these conditions are necessary and sufficient for optimality.
c. Suppose that the given point $\bar{x}$ happens to be a KKT point for $P$. Is $\bar{x}$ also a KKT point for $\bar{P}$? If so, why? If not, under what additional conditions can you make this claim?

d. Determine a closed-form expression for the solution $\hat{x}$ to Problem $\bar{P}$.

[4.31] Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad Ax = b \\
& \quad x \geq 0.
\end{align*}
\]

Let $\bar{x} = (\bar{x}_B, \bar{x}_N)$ be an extreme point, where $\bar{x}_B = B^{-1}b > 0$, $\bar{x}_N = 0$, and $A = [B, N]$ with $B$ invertible. Now, consider the following direction-finding problem:

\[
\begin{align*}
\text{Minimize} & \quad [\nabla_N f(\bar{x}) - \nabla_B f(\bar{x})B^{-1}N']d_N \\
\text{subject to} & \quad 0 \leq d_j \leq 1 \quad \text{for each nonbasic component } j,
\end{align*}
\]

where $\nabla_B f(\bar{x})$ and $\nabla_N f(\bar{x})$ denote the gradient of $f$ with respect to the basic and nonbasic variables, respectively. Let $d_N$ be an optimal solution, and let $d_B = -B^{-1}Nd_N$. Show that if $d' = (d'_B, d'_N) \neq (0, 0)$, it is an improving feasible direction. What are the implications of $d = 0$?

[4.32] Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad \sum_{j=1}^{n} f_j(x_j) \\
\text{subject to} & \quad \sum_{j=1}^{n} x_j = 1 \\
& \quad x_j \geq 0 \quad \text{for } j = 1, \ldots, n.
\end{align*}
\]

Suppose that $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_n)^t \geq 0$ solves the problem. Letting $\delta_j = \partial f_j(\bar{x})/\partial x_j$, show that there exists a scalar $k$ such that

$\delta_j \geq k$ \quad \text{and} \quad (\delta_j - k)\bar{x}_j = 0 \quad \text{for } j = 1, \ldots, n.$

[4.33] Let $c$ be an $n$-vector, $b$ an $m$-vector, $A$ an $m \times n$ matrix, and $H$ a symmetric $n \times n$ positive definite matrix. Consider the following two problems:
• Minimize \( c'x + \frac{1}{2}x'Hx \)
subject to \( Ax \leq b \);

• Minimize \( h'v + \frac{1}{2}v'Gv \)
subject to \( v \geq 0 \),

where \( G = AH^{-1}A' \) and \( h = AH^{-1}c + b \). Investigate the relationship between the KKT conditions of these two problems.

[4.34] Consider the following problem:

Minimize \(-x_1 + x_2\)
subject to \( x_1^2 + x_2^2 - 2x_1 = 0 \)\( \quad (x_1, x_2) \in X \),

where \( X \) is the convex combinations of the points \((-1, 0), (0, 1), (1, 0), \) and \((0, -1)\).

a. Find the optimal solution graphically.
b. Do the Fritz John or KKT conditions hold at the optimal solution in Part a? If not, explain in terms of Theorems 4.3.2 and 4.3.7.
c. Replace the set \( X \) by a suitable system of inequalities and answer Part b. What are your conclusions?

[4.35] Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \), where \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) and \( g_i: \mathbb{R}^n \rightarrow \mathbb{R} \) for \( i = 1, \ldots, m \), are all differentiable functions. We know that if \( \bar{x} \) is a local minimum, then \( F \cap D = \emptyset \), where \( F \) and \( D \) are, respectively, the set of improving and feasible directions. Show, giving examples, that the converse is false even if \( f \) is convex or if the feasible region is convex (although not both). However, suppose that there exists an \( \varepsilon \)-neighborhood \( N_\varepsilon(\bar{x}) \) about \( \bar{x} \), \( \varepsilon > 0 \), such that \( f \) is pseudoconvex and \( g_i \) for \( i \in I = \{i : g_i(\bar{x}) = 0\} \) are quasiconvex over \( N_\varepsilon(\bar{x}) \). Show that \( \bar{x} \) is a local minimum if and only if \( F \cap D = \emptyset \). (Hint: Examine Lemma 4.2.3 and Theorem 4.2.5.) Extend this result to include equality constraints.

[4.36] Consider the following problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) and \( h_\ell(x) = 0 \) for \( i = 1, \ldots, \ell \). Suppose that \( \bar{x} \) solves the problem locally, and let \( I = \{i : g_i(\bar{x}) = 0\} \). Furthermore, suppose that each \( g_i \) for \( i \in I \) is differentiable at \( \bar{x} \), each \( g_i \) for \( i \not\in I \) is continuous at \( \bar{x} \), and \( h_1, \ldots, h_\ell \) are affine; that is, each \( h_\ell \) is of the form \( h_\ell(x) = a_\ell'x - b_\ell \).

a. Show that \( F_0 \cap G \cap H_0 = \emptyset \), where
\[
F_0 = \{d : \nabla f(\bar{x})'d < 0\} \]
\[ H_0 = \{ d : \nabla h_i(\bar{x})^t d = 0 \text{ for } i = 1, \ldots, \ell \} \]

\[ G = \{ d : \nabla g_i(\bar{x})^t d \leq 0 \text{ for } i \in J \text{ and } \nabla g_i(\bar{x})^t d < 0 \text{ for } i \in I - J \}, \]

where \( J = \{ i \in I : g_i \text{ is pseudoconcave at } \bar{x} \}. \]

b. Show how this condition can be verified by using linear programming.

c. Dropping the equality constraints \( h_i(\bar{x}) = 0, \ i = 1, \ldots, \ell, \) from the problem and letting \( D \) denote the resulting set of feasible directions, show that \( G \subseteq D \) and hence that \( F_0 \cap G = \emptyset. \)

[4.37] Consider the following problem:

Maximize \( x_1^2 + 4x_1x_2 + x_2^2 \)

subject to \( x_1^2 + x_2^2 = 1. \)

a. Using the KKT conditions, find an optimal solution to the problem.

b. Test for the second-order optimality conditions.

c. Does the problem have a unique optimal solution?

[4.38] Consider the following problem:

Maximize \( 3x_1 - x_2 + x_2^3 \)

subject to \( x_1 + x_2 + x_3 \leq 0 \)

\( -x_1 + 2x_2 + x_3^2 = 0. \)

a. Write the KKT optimality conditions.

b. Test for the second-order optimality conditions.

c. Argue why this problem is unbounded.

[4.39] Consider the following problem:

Maximize \( (x_1 - 2)^2 + (x_2 - 3)^2 \)

subject to \( 3x_1 + 2x_2 \geq 6 \)

\( -x_1 + x_2 \leq 3 \)

\( x_1 \leq 2. \)

a. Graphically, find all locally maximizing solutions. What is the global maximum for this problem?

b. Repeat Part a analytically, using first- and second-order KKT optimality conditions along with any other formal optimality characterizations.

[4.40] Consider the Problem of Example 4.4.4 for the case \( k = 1. \) Provide an analytical argument to show that \( \bar{x} = (0,0)^t \) is an optimal solution. By examining a sequence of values of \( k \rightarrow 1^- \) with respect to the point \((0, 0),\) explain why the second-order optimality conditions are unable to resolve this case.
Consider the problem to minimize \( f(x) \) subject to \( Ax \leq b \). Suppose that \( \bar{x} \) is a feasible solution such that \( A_1 \bar{x} = b_1 \) and \( A_2 \bar{x} < b_2 \), where \( A^t = (A_1^t, A_2^t) \) and \( b^t = (b_1^t, b_2^t) \). Assuming that \( A_1 \) has full rank, the matrix \( P \) that projects any vector unto the nullspace of \( A_1 \) is given by
\[
P = I - A_1^t (A_1 A_1^t)^{-1} A_1.
\]

a. Let \( \tilde{d} = -PVf(\bar{x}) \). Show that if \( \tilde{d} \neq 0 \), it is an improving feasible direction; that is, \( \bar{x} + \lambda \tilde{d} \) is feasible and \( f(\bar{x} + \lambda \tilde{d}) < f(\bar{x}) \) for \( \lambda > 0 \) and sufficiently small.

b. Suppose that \( \tilde{d} = 0 \) and that \( u = -(A_1 A_1^t)^{-1} A_1 Vf(\bar{x}) \geq 0 \). Show that \( \bar{x} \) is a KKT point.

c. Show that \( \tilde{d} \) generated above is of the form \( \lambda \hat{d} \) for some \( \lambda > 0 \), where \( \hat{d} \) is an optimal solution to the following problem:

\[
\text{Minimize } Vf(\bar{x})' d \\
\text{subject to } A_1 d = 0 \\
\|d\|^2 \leq 1.
\]

d. Make all possible simplifications if \( A = -I \) and \( b = 0 \), that is, if the constraints are of the form \( x \geq 0 \).

Consider the following problem:
\[
\text{Minimize } x_1^2 - x_1 x_2 + 2 x_2^2 - 4 x_1 - 5 x_2 \\
\text{subject to } x_1 + 2 x_2 \leq 6 \\
x_1 \leq 2 \\
x_1, \ x_2 \geq 0.
\]

a. Solve the problem geometrically and verify the optimality of the solution obtained by the KKT conditions.

b. Find the direction \( \tilde{d} \) of Exercise 4.41 at the optimal solution. Verify that \( \tilde{d} = 0 \) and that \( u \geq 0 \).

c. Find the direction \( \tilde{d} \) of Exercise 4.41 at \( \bar{x} = (1, 5/2)^t \). Verify that \( \tilde{d} \) is an improving feasible direction. Also, verify that the optimal solution \( \tilde{d} \) of Part c of Exercise 4.41 indeed points along \( \tilde{d} \).

Let \( A \) be an \( m \times n \) matrix of rank \( m \), and let \( P = I - A' (AA')^{-1} A \) be the matrix that projects any vector onto the null space of \( A \). Define \( C = \{d : Ad = 0\} \), and let \( H \) be an \( n \times n \) symmetric matrix. Show that \( d \in C \) if and only if \( d =\)
Pw for some \( w \in \mathbb{R}^n \). Show that \( d'HD \geq 0 \) for all \( d \in C \) if and only if \( P'HP \) is positive semidefinite.

[4.44] Consider Problem P to minimize \( f(x) \) subject to \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \). Let \( \bar{x} \) be a feasible solution and define \( A \) as an \( \ell \times n \) matrix whose rows represent \( \nabla h_i(\bar{x})' \) for \( i = 1, \ldots, \ell \) and assume that \( A \) is of rank \( \ell \). Define \( P = I - A'(AA')^{-1}A \) as in Exercise 4.41 to be the matrix that projects any vector onto the nullspace of \( A \). Explain how Exercise 4.43 relates to checking the second-order necessary conditions for Problem P. How can you extend this to checking second-order sufficient conditions? Illustrate using Example 4.4.4.

[4.45] Consider the problem to maximize \( 3x_1x_2 + 2x_2x_3 + 12x_1x_3 \) subject to \( 6x_1 + x_2 + 4x_3 = 6 \). Using first- and second-order KKT optimality conditions, show that \( \bar{x} = (1/3, 2, 1/2)' \) is a local maximum. Use Exercise 4.44 to check the second-order sufficient conditions.

[4.46] Consider the following problem:

Minimize \( c'x + \frac{1}{2}x'HX \)

subject to \( Ax \leq b \),

where \( c \) is an \( n \)-vector, \( b \) is an \( m \)-vector, \( A \) is an \( m \times n \) matrix, and \( H \) is an \( n \times n \) symmetric matrix.

a. Write the second-order necessary optimality conditions of Theorem 4.4.3. Make all possible simplifications.

b. Is it necessarily true that every local minimum to the above problem is also a global minimum? Prove, or give a counter example.

c. Provide the first- and second-order necessary optimality conditions for the special case where \( c = 0 \) and \( H = I \). In this case the problem reduces to finding the point in a polyhedral set closest to the origin. (The above problem is referred to in the literature as a least distance programming problem.)

[4.47] Investigate the relationship between the optimal solutions and the KKT conditions for the following two problems, where \( \lambda \geq 0 \) is a given fixed vector.

\[
\begin{align*}
P & : \text{Minimize } f(x) \text{ subject to } x \in X, \ g(x) \leq 0. \\
P' & : \text{Minimize } f(x) \text{ subject to } x \in X, \ A'g(x) \leq 0.
\end{align*}
\]

(Problem P' has only one constraint and is referred to as a surrogate relaxation of Problem P.)
Consider Problem P, to minimize \( f(x) \) subject to \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \), where \( f: \mathbb{R}^n \to \mathbb{R} \) and \( h_i: \mathbb{R}^n \to \mathbb{R} \) for \( i = 1, \ldots, \ell \) are all continuously differentiable functions. Let \( \bar{x} \) be a feasible solution and define an \( \ell \times \ell \) Jacobian submatrix \( J \) as

\[
J = \begin{bmatrix}
\frac{\partial h_1(x)}{\partial x_1} & \frac{\partial h_1(x)}{\partial x_\ell} \\
\vdots & \vdots \\
\frac{\partial h_\ell(x)}{\partial x_1} & \frac{\partial h_\ell(x)}{\partial x_\ell}
\end{bmatrix}.
\]

Assume that \( J \) is nonsingular, so, in particular, \( \nabla h_i(\bar{x}), \ i = 1, \ldots, \ell \), are linearly independent. Under these conditions, the Implicit Function Theorem (see Exercise 4.49) asserts that if we define \( y = (x_{\ell+1}, \ldots, x_n)' \in \mathbb{R}^{n-\ell} \) with \( \bar{y} = (\bar{x}_{\ell+1}, \ldots, \bar{x}_n)' \), there exists a neighborhood of \( \bar{y} \) over which the (first) \( \ell \) variables \( x_{\ell+1}, \ldots, x_n \) can (implicitly) be solved for in terms of the \( y \) variables using the \( \ell \) equality constraints. More precisely, there exists a neighborhood of \( \bar{y} \) and a set of functions \( \psi_1(y), \ldots, \psi_\ell(y) \) such that over this neighborhood, we have that \( \psi_1(y), \ldots, \psi_\ell(y) \) are continuously differentiable, \( \psi_i(\bar{y}) = \bar{x}_i \) for \( i = 1, \ldots, \ell \), and \( h_i[\psi_1(y), \ldots, \psi_\ell(y), y] = 0 \) for \( i = 1, \ldots, \ell \).

Now suppose that \( \bar{x} \) is a local minimum and that the above assumptions hold true. Argue that \( \bar{y} \) must be a local minimum for the unconstrained function \( F(y) = f[\psi_1(y), \ldots, \psi_\ell(y), y]: \mathbb{R}^{n-\ell} \to \mathbb{R} \). Using the first-order necessary optimality conditions \( \nabla F(\bar{y}) = 0 \) for unconstrained problems, derive the KKT necessary optimality conditions for Problem P. In particular, show that the Lagrangian multiplier vector \( \bar{\nu} \) in the dual feasibility condition \( \nabla f(\bar{x}) + [\nabla h(\bar{x})]' \bar{\nu} \), where \( \nabla h(\bar{x}) \) is the matrix whose rows are \( \nabla h_1(\bar{x})', \ldots, \nabla h_\ell(\bar{x})' \), is given uniquely by

\[
\bar{\nu} = -J^{-1} \left[ \frac{\partial f(\bar{x})}{\partial x_1}, \ldots, \frac{\partial f(\bar{x})}{\partial x_\ell} \right]'^\ell.
\]

Consider Problem P, to minimize \( f(x) \) subject to \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \), where \( x \in \mathbb{R}^n \) and where all objective and constraint functions are continuously differentiable. Suppose that \( \bar{x} \) is a local minimum for P and that the gradients \( \nabla h_i(\bar{x}), \ i = 1, \ldots, \ell \), are linearly independent. Use the implicit function theorem stated below to derive the KKT optimality conditions for P. Extend this to
include inequality constraints \( g_i(x) \leq 0, \ i = 1, \ldots, m \) as well. (Hint: See Exercise 4.48.)

**Implicit Function Theorem (see Taylor and Mann [1983], for example).** Suppose that \( \phi_i(x), \ i = 1, \ldots, p \) (representing the binding constraints at \( \bar{x} \)), are continuously differentiable functions, and suppose that the gradients \( \nabla \phi_i(\bar{x}), \ i = 1, \ldots, p, \) are linearly independent, where \( p < n \). Denote \( \phi = (\phi_1, \ldots, \phi_p) : \mathbb{R}^n \rightarrow \mathbb{R}^p. \) Hence, \( \phi(\bar{x}) = 0 \) and we can partition \( x' = (x'_B, x'_N) \), where \( x_B \in \mathbb{R}^p \) and \( x_N \in \mathbb{R}^{n-p} \) such that for the corresponding partition \( [\nabla_B \phi(x), \nabla_N \phi(x)] \) of the Jacobian \( \nabla \phi(x) \), the \( p \times p \) submatrix \( \nabla_B \phi(\bar{x}) \) is nonsingular. Then the following holds true: There exists an open neighborhood \( N_\epsilon(\bar{x}) \subseteq \mathbb{R}^n, \ \epsilon > 0, \) an open neighborhood \( N'_\epsilon(\bar{x}_N) \subseteq \mathbb{R}^{n-p}, \ \epsilon' > 0, \) and a function \( \psi : \mathbb{R}^{n-p} \rightarrow \mathbb{R}^p \) that is continuously differentiable on \( N'_\epsilon(\bar{x}_N) \) such that

1. \( \bar{x}_B = \psi(\bar{x}_N). \)
2. For every \( x_N \in N'_\epsilon(\bar{x}_N) \), we have \( \phi[\psi(x_N), x_N] = 0. \)
3. The Jacobian \( \nabla \phi(x) \) has full row rank \( p \) for each \( x \in N_\epsilon(\bar{x}). \)
4. For any \( x_N \in N'_\epsilon(\bar{x}_N) \), the Jacobian \( \nabla \psi(x_N) \) is given by the (unique) solution to the linear equation system

\[
[\nabla_B \phi(\nabla \psi(x_N), x_N)] \nabla \psi(x_N) = -\nabla_N \phi(\psi(x_N), x_N).
\]

4.50 A differentiable function \( \psi : \mathbb{R}^n \rightarrow \mathbb{R} \) is said to be an \( \eta \)-invex function if there exists some (arbitrary) function \( \eta : \mathbb{R}^{2n} \rightarrow \mathbb{R}^n \) such that for each \( x_1, x_2 \in \mathbb{R}^n, \ \psi(x_2) \geq \psi(x_1) + \nabla \psi(x_1)' \eta(x_1, x_2). \) Furthermore, \( \psi \) is said to be an \( \eta \)-pseudoinvex function if \( \nabla \psi(x_1)' \eta(x_1, x_2) \geq 0 \) implies that \( \psi(x_2) \geq \psi(x_1). \) Similarly, \( \psi \) is said to be an \( \eta \)-quasi-invex function if \( \psi(x_2) \leq \psi(x_1) \) implies that \( \nabla \psi(x_1)' \eta(x_1, x_2) \leq 0. \)

a. When invex is replaced by convex in the usual sense, what is \( \eta(x_1, x_2) \) defined to be?

b. Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) where \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) and \( g_i : \mathbb{R}^n \rightarrow \mathbb{R} \) for \( i = 1, \ldots, m \) are all differentiable functions. Let \( \bar{x} \) be a KKT point. Show that \( \bar{x} \) is optimal if \( f \) and \( g_i \) for \( i \in I = \{i : g_i(\bar{x}) = 0\} \) are all \( \eta \)-invex.

c. Repeat Part b if \( f \) is \( \eta \)-pseudoinvex and \( g_i, \ i \in I, \) are \( \eta \)-quasi-invex.

(The reader is referred to Hanson [1981] and to Hanson and Mond [1982, 1987] for discussions on invex functions and their uses.)
Notes and References

In this chapter we begin by developing first- and second-order optimality conditions for unconstrained optimization problems in Section 4.1. These classical results can be found in most textbooks dealing with real analysis. For more details on this subject relating to higher-order necessary and sufficiency conditions, refer to Gue and Thomas [1968] and Hancock [1960]; and for information regarding the handling of equality constraints via the Lagrangian multiplier rule, refer to Bartle [1976] and Rudin [1964].

In Section 4.2 we treat the problem of minimizing a function in the presence of inequality constraints and develop the Fritz John [1948] necessary optimality conditions. A weaker form of these conditions, in which the nonnegativity of the multipliers was not asserted, was derived by Karush [1939]. Under a suitable constraint qualification, the Lagrangian multiplier associated with the objective function is positive, and the Fritz John conditions reduce to those of Kuhn and Tucker [1951], which were derived independently. Even though the latter conditions were originally derived by Karush [1939] using calculus of variations, this work had not received much attention, since it was not published. However, we refer to these conditions as KKT conditions, recognizing Karush, Kuhn, and Tucker. An excellent historical review of optimality conditions for nonlinear programming can be found in Kuhn [1976] and Lenstra et al. [1991]. Kyparisis [1985] presents a necessary and sufficient condition for the KKT Lagrangian multipliers to be unique. Gehner [1974] extends the FJ optimality conditions to the case of semi-infinite nonlinear programming problems, where there are an infinite number of parametrically described equality and inequality constraints. The reader may refer to the following references for further study of the Fritz John and KKT conditions: Abadie [1967b], Avriel [1967], Canon et al. [1966], Gould and Tolle [1972], Luenberger [1973], Mangasarian [1969a], and Zangwill [1969].

Mangasarian and Fromovitz [1967] generalized the Fritz John conditions for handling both equality and inequality constraints. Their approach used the implicit function theorem. In Section 4.3 we develop the Fritz John conditions for equality and inequality constraints by constructing a feasible arc, as in the work of Fiacco and McCormick [1968].

In Sections 4.2 and 4.3 we show that the KKT conditions are indeed sufficient for optimality under suitable convexity assumptions. This result was proved by Kuhn and Tucker [1951] if the functions $f_i, g_i$ for $i \in I$ are convex, the functions $h_i$ for all $i$ are affine, and the set $X$ is convex. This result was generalized later, so that weaker convexity assumptions are needed to guarantee optimality, as shown in Sections 4.2 and 4.3 (see Mangasarian [1969a]). The reader may also refer to Bhatt and Misra [1975], who relaxed the condition that $h_i$ be affine, provided that the associated Lagrangian multiplier has the correct sign. Further generalizations using invex functions can be found in Hanson [1981] and Hanson and Mond [1982].

Other generalizations and extensions of the Fritz John and KKT conditions were developed by many authors. One such extension is to relax the condition that the set $X$ is open. In this case we obtain necessary optimality
conditions of the minimum principle type. For details on this type of optimality conditions, see Bazaraa and Goode [1972], Canon et al. [1970], and Mangasarian [1969a]. Another extension is to treat the problem in an infinite-dimensional setting. The interested reader may refer to Canon et al. [1970], Dubovitskii and Milyutin [1965], Gehner [1974], Guignard [1969], Halkin and Neustadt [1966], Hestenes [1966], Neustadt [1969], and Varaiya [1967].

In Section 4.4 we address second-order necessary and sufficient optimality conditions for constrained problems, developed initially by McCormick [1967]. Our second-order necessary optimality condition is stronger than that presented by McCormick [1967] (see Fletcher [1987] and Ben-Tal [1980]). For a discussion on checking these conditions based on eigenvalues computed over a projected tangential subspace, or based on bordered Hessian matrices, we refer the reader to Luenberger [1973a/1984]. See Exercise 4.44 for a related approach. For extensions and additional study of this topic, we refer the reader to Avriel [1976], Baccari and Trad [2004], Ben-Tal [1980], Ben-Tal and Zowe [1982], Fletcher [1983], Luenberger [1973a/1984], McCormick [1967], and Messerli and Polak [1969].
Chapter 5

Constraint Qualifications

In Chapter 4 we considered Problem P to minimize \( f(x) \) subject to \( x \in X \) and \( g_i(x) \leq 0, i = 1, \ldots, m \). We obtained the Karush-Kuhn-Tucker (KKT) necessary conditions for optimality by deriving the Fritz John conditions and then asserting that the multiplier associated with the objective function is positive at a local optimum when a constraint qualification is satisfied. In this chapter we develop the KKT conditions directly without first deriving the Fritz John conditions. This is done under various constraint qualifications for problems having inequality constraints and for problems having both inequality and equality constraints.

Following is an outline of the chapter.

**Section 5.1: Cone of Tangents**
We introduce the cone of tangents \( T \) and show that \( F_0 \cap T = \emptyset \) is a necessary condition for local optimality. Using a constraint qualification, we derive the KKT conditions directly for problems having inequality constraints.

**Section 5.2: Other Constraint Qualifications**
We introduce other cones contained in the cone of tangents. Making use of these cones, we present various constraint qualifications that validate the KKT conditions. Relationships among these constraint qualifications are explored.

**Section 5.3: Problems Having Inequality and Equality Constraints**
The results of Section 5.2 are extended to problems having equality and inequality constraints.

### 5.1 Cone of Tangents

In Section 4.2 we discussed the KKT necessary optimality conditions for problems having inequality constraints. In particular, we showed that local optimality implies that \( F_0 \cap G_0 = \emptyset \), which in turn implies the Fritz John conditions. Under the linear independence constraint qualification, or, more generally, under the constraint qualification \( G_0 \neq \emptyset \), we deduced that the Fritz John conditions can only be satisfied if the Lagrangian multiplier associated with the objective function is positive. This led to the KKT conditions. This process is summarized in the following flowchart.
In this section we derive the KKT conditions directly without first obtaining the Fritz John conditions. As shown in Theorem 5.1.2, a necessary condition for local optimality is that \( F_0 \cap T = \emptyset \), where \( T \) is the cone of tangents introduced in Definition 5.1.1. Using the constraint qualification \( T = G' \), where \( G' \) is as defined in Theorem 5.1.3 (see also Theorem 4.2.15), we get \( F_0 \cap G' = \emptyset \). Using Farkas's theorem, this statement gives the KKT conditions. This process is summarized in the following flowchart.

**5.1.1 Definition**

Let \( S \) be a nonempty set in \( \mathbb{R}^n \), and let \( \bar{x} \in \text{cl } S \). The *cone of tangents* of \( S \) at \( \bar{x} \), denoted by \( T \), is the set of all directions \( d \) such that \( d = \lim_{k \to \infty} \lambda_k (x_k - \bar{x}) \), where \( \lambda_k > 0 \), \( x_k \in S \) for each \( k \), and \( x_k \to \bar{x} \).

From the above definition, it is clear that \( d \) belongs to the cone of tangents if there is a feasible sequence \( \{x_k\} \) converging to \( \bar{x} \) such that the directions \( x_k - \bar{x} \) converge to \( d \). Exercise 5.1 provides alternative equivalent descriptions for the cone of tangents \( T \); and in Exercise 5.2, we ask the reader to show that the cone of tangents is indeed a closed cone. Figure 5.1 illustrates some examples of the cone of tangents, where the origin is translated to \( \bar{x} \) for convenience.

Theorem 5.1.2 shows that for a problem of the form to minimize \( f(x) \) subject to \( x \in S, F_0 \cap T = \emptyset \) is, indeed, a necessary condition for optimality. Later we specify \( S \) to be the set \( \{x \in X : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m\} \).

**5.1.2 Theorem**

Let \( S \) be a nonempty set in \( \mathbb{R}^n \), and let \( \bar{x} \in S \). Furthermore, suppose that \( f: \mathbb{R}^n \to \mathbb{R} \) is differentiable at \( \bar{x} \). If \( \bar{x} \) locally solves the problem to minimize \( f(x) \)
subject to $x \in S$, $F_0 \cap T = \emptyset$, where $F_0 = \{d : \nabla f(\bar{x})'d < 0\}$ and $T$ is the cone of tangents of $S$ at $\bar{x}$.

**Proof**

Let $d \in T$, that is, $d = \lim_{k \to \infty} \lambda_k (x_k - \bar{x})$, where $\lambda_k > 0$, $x_k \in S$ for each $k$, and $x_k \to \bar{x}$. By the differentiability of $f$ at $\bar{x}$, we get

$$f(x_k) - f(\bar{x}) = \nabla f(\bar{x})'(x_k - \bar{x}) + \|x_k - \bar{x}\| \alpha(\bar{x}; x_k - \bar{x}),$$

where $\alpha(\bar{x}; x_k - \bar{x}) \to 0$ as $x_k \to \bar{x}$. Noting the local optimality of $\bar{x}$, we have for $k$ large enough that $f(x_k) \geq f(\bar{x})$; so from (5.1) we get

$$\nabla f(\bar{x})'(x_k - \bar{x}) + \|x_k - \bar{x}\| \alpha(\bar{x}; x_k - \bar{x}) \geq 0.$$

Multiplying by $\lambda_k > 0$ and taking the limit as $k \to \infty$, the above inequality implies that $\nabla f(\bar{x})'d \geq 0$. Hence, we have shown that $d \in T$ implies that $\nabla f(\bar{x})'d \geq 0$ and therefore that $F_0 \cap T = \emptyset$. This completes the proof.

It is worth noting that the condition $F_0 \cap T = \emptyset$ does not necessarily imply that $\bar{x}$ is a local minimum. Indeed, this condition will hold true whenever $F_0 = \emptyset$, for example, which we know is not sufficient for local optimality. However, if there exists an $\varepsilon$-neighborhood $N_\varepsilon(\bar{x})$ about $\bar{x}$ such that $N_\varepsilon(\bar{x}) \cap S$ is convex and $f$ is pseudoconvex over $N_\varepsilon(\bar{x}) \cap S$, $F_0 \cap T = \emptyset$ is sufficient to claim that $\bar{x}$ is a local minimum (see Exercise 5.3).

**Abadie Constraint Qualification**

In Theorem 5.1.3, we derive the KKT conditions under the constraint qualification $T = G'$, which is credited to Abadie.
5.1.3 Theorem (Karush–Kuhn–Tucker Necessary Conditions)

Let $X$ be a nonempty set in $R^n$, and let $f: R^n \to R$ and $g_i: R^n \to R$ for $i = 1, \ldots, m$. Consider the problem: Minimize $f(x)$ subject to $x \in X$ and $g_i(x) \leq 0$ for $i = 1, \ldots, m$. Let $\bar{x}$ be a feasible solution, and let $I = \{i : g_i(\bar{x}) = 0\}$. Suppose that $f$ and $g_i$ for $i \in I$ are differentiable at $\bar{x}$. Furthermore, suppose that the constraint qualification $T = G'$ holds true, where $T$ is the cone of tangents of the feasible region at $\bar{x}$ and $G' = \{d : \nabla g_i(\bar{x})^t d \leq 0$ for $i \in I\}$. If $\bar{x}$ is a local optimal solution, there exist nonnegative scalars $u_i$ for $i \in I$ such that

$$\nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) = 0.$$

**Proof**

By Theorem 5.1.2 we have that $F_0 \cap T = \emptyset$, where $F_0 = \{d : \nabla f(\bar{x})^t d < 0\}$. By assumption, $T = G'$, so that $F_0 \cap G' = \emptyset$. In other words, the following system has no solution:

$$\nabla f(\bar{x})^t d < 0, \quad \nabla g_i(\bar{x})^t d \leq 0 \quad \text{for } i \in I.$$

Hence, by Theorem 2.4.5 (Farkas's theorem), the result follows (see also Theorem 4.2.15).

The reader may verify that in Example 4.2.10, the constraint qualification $T = G'$ does not hold true at $\bar{x} = (1,0)^t$. Note that the Abadie constraint qualification $T = G'$ could be stated equivalently as $T \supseteq G'$, since $T \subseteq G'$ is always true (see Exercise 5.4). Note that openness of the set $X$ and continuity of $g_i$ at $\bar{x}$ for $i \notin I$ were not assumed explicitly in Theorem 5.1.3. However, without these assumptions, it is unlikely that the constraint qualification $T \supseteq G'$ would hold true (see Exercise 5.5).

**Linearly Constrained Problems**

Lemma 5.1.4 shows that if the constraints are linear, the Abadie constraint qualification is satisfied automatically. This also implies that the KKT conditions are always necessary for problems having linear constraints, whether the objective function is linear or nonlinear. As an alternative proof that does not employ the cone of tangents, note that if $\bar{x}$ is a local minimum, $F_0 \cap D = \emptyset$. Now, by Lemma 4.2.4, if the constraints are linear, $D = G' = \{d \neq 0 : \nabla g_i(\bar{x})^t d \leq 0$, for each $i \in I\}$. Hence, $F_0 \cap D = \emptyset \iff F_0 \cap G' = \emptyset$, which holds true if and only if $\bar{x}$ is a KKT point by Theorem 4.2.15.
5.1.4 Lemma

Let $A$ be an $m \times n$ matrix, let $b$ be an $m$-vector, and let $S = \{x : Ax \leq b\}$. Suppose that $\bar{x} \in S$ is such that $A_1\bar{x} = b_1$ and $A_2\bar{x} < b_2$, where $A' = (A_1', A_2')$ and $b' = (b_1', b_2')$. Then $T = G'$, where $T$ is the cone of tangents of $S$ at $\bar{x}$ and $G' = \{d : A_1d \leq 0\}$.

Proof

If $A_1$ is vacuous, then $G' = \mathbb{R}^n$. Furthermore, $\bar{x} \in \text{int} S$ and hence $T = \mathbb{R}^n$. Thus, $G' = T$. Now, suppose that $A_1$ is not vacuous. Let $d \in T$; that is, $d = \lim_{k \to \infty} \lambda_k (x_k - \bar{x})$, where $x_k \in S$ and $\lambda_k > 0$ for each $k$. Then

$$A_1(x_k - \bar{x}) \leq b_1 - b_1 = 0. \quad (5.2)$$

Multiplying (5.2) by $\lambda_k > 0$ and taking the limit as $k \to \infty$, it follows that $A_1d \leq 0$. Thus, $d \in G'$ and $T \subseteq G'$. Now let $d \in G'$; that is, $A_1d \leq 0$. We need to show that $d \in T$. Since $A_2\bar{x} < b_2$, there is a $\delta > 0$ such that $A_2(\bar{x} + \lambda d) < b_2$ for all $\lambda \in (0, \delta)$. Furthermore, since $A_1\bar{x} = b_1$ and $A_1d \leq 0$, then $A_1(\bar{x} + \lambda d) \leq b_1$ for all $\lambda > 0$. Therefore, $\bar{x} + \lambda d \in S$ for each $\lambda \in (0, \delta)$. This shows automatically that $d \in T$. Therefore, $T = G'$, and the proof is complete.

5.2 Other Constraint Qualifications

The KKT conditions have been developed by many authors under various constraint qualifications. In this section we present some of the more important constraint qualifications. In Section 5.1 we learned that local optimality implies that $F_0 \cap T = \emptyset$ and that the KKT conditions follow under the constraint qualification $T = G'$. If we define a cone $C \subseteq T$, $F_0 \cap T = \emptyset$ also implies that $F_0 \cap C = \emptyset$. Therefore, any constraint qualification of the form $C = G'$ will lead to the KKT conditions. In fact, since $C \subseteq T \subseteq G'$, the constraint qualification $C = G'$ implies that $T = G'$ and is therefore more restrictive than Abadie’s constraint qualification. This process is illustrated in the following flowchart:
We present below several such cones whose closures are contained in $T$. Here the feasible region $S$ is given by $\{x \in X : g_i(x) \leq 0, \ i = 1, \ldots, m\}$. The vector $\bar{x}$ is a feasible point, and $I = \{i : g_i(\bar{x}) = 0\}$.

**Cone of Feasible Directions of $S$ at $\bar{x}$**

This cone was introduced in Definition 4.2.1. The cone of feasible directions, denoted by $D$, is the set of all nonzero vectors $d$ such that $\bar{x} + \lambda d \in S$ for $\lambda \in (0, \delta)$ for some $\delta > 0$.

**Cone of Attainable Directions of $S$ at $\bar{x}$**

A nonzero vector $d$ belongs to the cone of attainable directions, denoted by $A$, if there exist a $\delta > 0$ and an $\alpha : R \to R^n$ such that $\alpha(\lambda) \in S$ for $\lambda \in (0, \delta)$, $\alpha(0) = \bar{x}$, and

$$\lim_{\lambda \to 0^+} \frac{\alpha(\lambda) - \alpha(0)}{\lambda} = d.$$

In other words, $d$ belongs to the cone of attainable directions if there is a feasible arc starting from $\bar{x}$ that is tangential to $d$.

**Cone of Interior Directions of $S$ at $\bar{x}$**

This cone, denoted by $G_0$, was introduced in Section 4.2 and is defined as $G_0 = \{d : \nabla g_i(\bar{x})'d < 0 \text{ for } i \in I\}$. Note that if $X$ is open and each $g_i$ for $i \notin I$ is continuous at $\bar{x}$, then $d \in G_0$ implies that $\bar{x} + \lambda d$ belongs to the interior of the feasible region for $\lambda > 0$ and sufficiently small.

Lemma 5.2.1 shows that all the above cones and their closures are contained in $T$.

**5.2.1 Lemma**

Let $X$ be a nonempty set in $R^n$, and let $f : R^n \to R$ and $g_i : R^n \to R$ for $i = 1, \ldots, m$. Consider the problem to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$ and $x \in X$. Let $\bar{x}$ be a feasible point, and let $I = \{i : g_i(\bar{x}) = 0\}$. Suppose that each $g_i$ for $i \in I$ is differentiable at $\bar{x}$, and let $G' = \{d : \nabla g_i(\bar{x})'d \leq 0 \text{ for } i \in I\}$. Then

$$\text{cl } D \subseteq \text{cl } A \subseteq T \subseteq G',$$

where $D$, $A$, and $T$ are, respectively, the cone of feasible directions, the cone of attainable directions, and the cone of tangents of the feasible region at $\bar{x}$. Furthermore, if $X$ is open and each $g_i$ for $i \notin I$ is continuous at $\bar{x}$, then $G_0 \subseteq D$, so that
Where \( G_0 \) is the cone of interior directions of the feasible region at \( \bar{x} \).

**Proof**

It can easily be verified that \( D \subseteq A \subseteq T \subseteq G' \) and that, since \( T \) is closed (see Exercise 5.2), \( \text{cl}\, D \subseteq \text{cl}\, A \subseteq T \subseteq G' \). Now note that \( G_0 \subseteq D \) by Lemma 4.2.4. Hence, the second part of the lemma follows.

To illustrate how each of the containments considered above can be strict, consider the following example. In Figure 4.9, note that \( G_0 = \emptyset = \text{cl}\, G_0 \), since there are no interior directions, whereas \( D = \text{cl}\, D = G' \) is defined by the feasible direction along the edge incident at \( \bar{x} \). In regard to the cone of interior directions \( G_0 \), note that whereas any \( d \in G_0 \) is a direction leading to interior feasible solutions, it is not true that any feasible direction that leads to interior points belongs to \( G_0 \). For example, consider Example 4.3.5, illustrated in Figure 4.12, with the equalities replaced by “less than or equal to” inequalities. The set \( G_0 = \emptyset \) at \( \bar{x} = (1,0)' \), whereas \( d = (-1,0)' \) leads to interior feasible points.

To show that \( \text{cl}\, D \) can be a strict subset of \( \text{cl}\, A \), consider the region defined by \( x_1 - x_2^2 \leq 0 \) and \( -x_1 + x_2^2 \leq 0 \). The set of feasible points lies on the parabola \( x_1 = x_2^2 \). At \( \bar{x} = (0,0)' \), for example, \( D = \emptyset = \text{cl}\, D \), whereas \( \text{cl}\, A = \{ d : d = \lambda (0,1)' \text{ or } d = \lambda (0,-1)' , \lambda \geq 0 \} = G' \).

The possibility that \( \text{cl}\, A \neq T \) is a little more subtle. Suppose that the feasible region \( S \) is itself the sequence \{ \((1/k,0)' , k = 1, 2, ... \) \} formed by the intersection of suitable constraints (written as suitable inequalities). For example, we might have \( S = \{(x_1,x_2) : x_2 = h(x_1) , x_2 = 0 , 0 \leq x_1 \leq 1 \text{, where } h(x_1) = x_1^3 \sin(\pi/x_1) \text{ if } x_1 \neq 0 \text{ and } h(x_1) = 0 \text{ if } x_1 = 0 \} \). Then \( A = \emptyset = \text{cl}\, A \), since there are no feasible arcs. However, by definition, \( T = \{ d : d = \lambda (1,0)' , \lambda \geq 0 \} \), and it is readily verified that \( T = G' \).

Finally, Figure 4.7 illustrates an instance where \( T \) is a strict subset of \( G' \). Here \( T = \{ d : d = \lambda (-1,0)' , \lambda \geq 0 \} \), while \( G' = \{ d : d = \lambda (-1,0)' , \text{ or } d = \lambda (1,0)' , \lambda \geq 0 \} \).

We now present some constraint qualifications that validate the KKT conditions and discuss their interrelationships.

**Slater’s Constraint Qualification**

The set \( X \) is open, each \( g_i \) for \( i \in I \) is pseudoconvex at \( \bar{x} \), for each \( g_i \) for \( i \notin I \) is continuous at \( \bar{x} \), and there is an \( x \in X \) such that \( g_i(x) < 0 \) for all \( i \in I \).
Linear Independence Constraint Qualification

The set $X$ is open, each $g_i$ for $i \in I$ is continuous at $\bar{x}$, and $\nabla g_i(\bar{x})$ for $i \in I$ are linearly independent.

Cottle's Constraint Qualification

The set $X$ is open and each $g_i$ for $i \in I$ is continuous at $\bar{x}$, and $\text{cl} \ G_0 = G'$.

Zangwill's Constraint Qualification

$$\text{cl} \ D = G'.$$

Kuhn–Tucker's Constraint Qualification

$$\text{cl} \ A = G'.$$

Validity of the Constraint Qualifications and Their Interrelationships

In Theorem 5.1.3 we showed that the KKT necessary optimality conditions are necessarily true under Abadie's constraint qualification $T = G'$. We demonstrate below that all the constraint qualifications discussed above imply that of Abadie and hence, each validate the KKT necessary conditions. From Lemma 5.2.1 it is clear that Cottle's constraint qualification implies that of Zangwill, which implies that of Kuhn and Tucker, which in turn implies Abadie's qualification. We now show that the first two constraint qualifications imply that of Cottle.

First, suppose that Slater's constraint qualification holds true. Then there is an $x \in X$ such that $g_i(x) < 0$ for $i \in I$. Since $g_i(x) < 0$ and $g_i(\bar{x}) = 0$, then, by the pseudoconvexity of $g_i$ at $\bar{x}$, it follows that $\nabla g_i(\bar{x})' (x - \bar{x}) < 0$. Thus, $d = x - \bar{x}$ belongs to $G_0$. Therefore, $G_0 \neq \emptyset$ and the reader can verify that $\text{cl} \ G_0 = G'$ and hence that Cottle's constraint qualification holds true. Now, suppose that the linear independence constraint qualification is satisfied. Then $\sum_{i \in I} u_i \nabla g_i(\bar{x}) = 0$ has no nonzero solution. By Theorem 2.4.9 it follows that there exists a vector $d$ such that $\nabla g_i(\bar{x})' d < 0$ for all $i \in I$. Thus, $G_0 \neq \emptyset$, and Cottle's constraint qualification holds true. The relationships among the foregoing constraint qualifications are illustrated in Figure 5.2.

In discussing Lemma 5.2.1, we gave various examples in which for each consecutive pair in the string of containments $\text{cl} \ G_0 \subset \text{cl} \ D \subset \text{cl} \ A \subset T \subset G'$, the containment was strict and the larger set was equal to $G'$. Hence, these examples also illustrate that the implications of Figure 5.2 in regard to these sets are one-way implications. Figure 5.2 accordingly illustrates for each constraint qualification an instance where it holds true, whereas the preceding constraint qualification that makes a more restrictive assumption does not hold true.
Needless to say, if a local minimum $\bar{x}$ is not a KKT point, as in Example 4.2.10 and illustrated in Figure 4.7, for instance, no constraint qualification can possibly hold true.

Finally, we remark that Cottle's constraint qualification is equivalent to requiring that $G(0) \neq \emptyset$ (see Exercise 5.6). Moreover, we have seen that Slater's constraint qualification and the linear independence constraint qualification both imply Cottle's constraint qualification. Hence, whenever these constraint qualifications hold true at a local minimum $\bar{x}$, then $\bar{x}$ is a Fritz John point, with the Lagrangian multiplier $u_0$ associated with the objective function necessarily positive. In contrast, we might have Zangwill's, or Kuhn–Tucker's, or Abadie's constraint qualifications holding true at a local minimum $\bar{x}$, while $u_0$ might possibly be zero in some solution to the Fritz John conditions. However, since these are valid constraint qualifications, in such a case we must also have $u_0 > 0$ in some solution to the Fritz John conditions.

### 5.3 Problems Having Inequality and Equality Constraints

In this section we consider problems having both inequality and equality constraints. In particular, consider the following problem:
Minimize \( f(x) \)
subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \)
\( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \)
\( x \in X. \)

By Theorem 5.1.2 a necessary optimality condition is \( F_0 \cap T = \emptyset \) at a local minimum \( \tilde{x} \). By imposing the constraint qualification \( T = G' \cap H_0 \), where \( H_0 = \{ d : \nabla h_i(\tilde{x})'d = 0 \text{ for } i = 1, \ldots, \ell \} \), this implies that \( F_0 \cap G' \cap H_0 = \emptyset \). By using Farkas's theorem [or see Equation (4.16)], the KKT conditions follow. Theorem 5.3.1 reiterates this. The process is summarized in the following flowchart:

5.3.1 Theorem (Karush–Kuhn–Tucker Conditions)

Let \( f: \mathbb{R}^n \to \mathbb{R}, \ g_i: \mathbb{R}^n \to \mathbb{R} \) for \( i = 1, \ldots, m \), and \( h_i: \mathbb{R}^n \to \mathbb{R} \) for \( i = 1, \ldots, \ell \), and let \( X \) be a nonempty set in \( \mathbb{R}^n \). Consider the following problem:

Minimize \( f(x) \)
subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \)
\( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \)
\( x \in X. \)

Let \( \bar{x} \) locally solve the problem, and let \( I = \{ i : g_i(\bar{x}) = 0 \} \). Suppose that \( f, g_i \) for \( i \in I \), and \( h_i \) for \( i = 1, \ldots, \ell \) are differentiable at \( \bar{x} \). Suppose that the constraint qualification \( T = G' \cap H_0 \) holds true, where \( T \) is the cone of tangents of the feasible region at \( \bar{x} \), and

\[
G' = \{ d : \nabla g_i(\bar{x})'d \leq 0 \text{ for } i \in I \},
\]
\[
H_0 = \{ d : \nabla h_i(\bar{x})'d = 0 \text{ for } i = 1, \ldots, \ell \}.
\]

Then \( \bar{x} \) is a KKT point; that is, there exist scalars \( u_i \geq 0 \) for \( i \in I \) and \( v_i \) for \( i = 1, \ldots, \ell \) such that
\[ \nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{x}) = 0. \]

**Proof**

Since \( \bar{x} \) solves the problem locally, by Theorem 5.1.2 we have that \( F_0 \cap T = \emptyset \). By the constraint qualification, we have \( F_0 \cap G' \cap H_0 = \emptyset \); that is, the system \( Ad \leq 0 \) and \( c'd > 0 \) has no solution, where the rows of \( A \) are given by \( \nabla g_i(\bar{x})' \) for \( i \in I \), \( \nabla h_i(\bar{x})' \) and \( -\nabla h_i(\bar{x})' \) for \( i = 1, \ldots, \ell \), and \( c = -\nabla f(\bar{x}) \). By Theorem 2.4.5, the system \( A'y = c \) and \( y \geq 0 \) has a solution. This implies that there exist nonnegative scalars \( u_i \) for \( i \in I \) and \( \alpha_i, \beta_i \) for \( i = 1, \ldots, \ell \), such that

\[ \nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} \alpha_i \nabla h_i(\bar{x}) - \sum_{i=1}^{\ell} \beta_i \nabla h_i(\bar{x}) = 0. \]

Letting \( v_i = \alpha_i - \beta_i \) for each \( i \), the result follows.

We now present several constraint qualifications that validate the KKT conditions. These qualifications use several cones that were defined earlier in the chapter. By replacing each equality constraint by two equivalent inequalities, the role played by \( G' \) in the preceding section is now played by the cone \( G' \cap H_0 \). The reader may note that Zangwill's constraint qualification is omitted here, since the cone of feasible directions is usually equal to the zero vector in the presence of nonlinear equality constraints.

**Slater's Constraint Qualification**

The set \( X \) is open, each \( g_i \) for \( i \in I \) is pseudoconvex at \( \bar{x} \), each \( g_i \) for \( i \not\in I \) is continuous at \( \bar{x} \), each \( h_i \) for \( i = 1, \ldots, \ell \) is quasiconvex, quasiconcave, and continuously differentiable at \( \bar{x} \), and \( \nabla h_i(\bar{x}) \) for \( i = 1, \ldots, \ell \) are linearly independent. Furthermore, there exists an \( x \in X \) such that \( g_i(x) < 0 \) for all \( i \in I \) and \( h_i(x) = 0 \) for all \( i = 1, \ldots, \ell \).

**Linear Independence Constraint Qualification**

The set \( X \) is open, each \( g_i \) for \( i \not\in I \) is continuous at \( \bar{x} \), \( \nabla g_i(\bar{x}) \) for \( i \in I \) and \( \nabla h_i(\bar{x}) \) for \( i = 1, \ldots, \ell \), are linearly independent, and each \( h_i \) for \( i = 1, \ldots, \ell \) is continuously differentiable at \( \bar{x} \).
Cottle's Constraint Qualification

The set $X$ is open, each $g_i$ for $i \not\in I$ is continuous at $\bar{x}$, each $h_i$ for $i = 1, \ldots, \ell$ is continuously differentiable at $\bar{x}$, and $\nabla h_i(\bar{x})$ for $i = 1, \ldots, \ell$ are linearly independent. Furthermore, $\text{cl} (G_0 \cap H_0) = G' \cap H_0$. [This is equivalent to the Mangasarian-Fromovitz constraint qualification, which requires $\nabla h_i(\bar{x})$, $i = 1, \ldots, \ell$, to be linearly independent and that $G_0 \cap H_0 \neq \emptyset$; see Exercise 5.7.]

Kuhn-Tucker's Constraint Qualification

$$\text{cl} A = G' \cap H_0.$$ 

Abadie's Constraint Qualification

$$T = G' \cap H_0.$$ 

Validity of the Constraint Qualifications and Their Interrelationships

In Theorem 5.3.1 we showed that the KKT conditions hold true if Abadie's constraint qualification $T = G' \cap H_0$ is satisfied. We demonstrate below that all the constraint qualifications given above imply that of Abadie, and hence, each validates the KKT necessary conditions.

As in Lemma 5.2.1, the reader can easily verify that $\text{cl} A \subseteq T \subseteq G' \cap H_0$. Now, suppose that $X$ is open, $g_i$ for each $i \not\in I$ is continuous at $\bar{x}$, $h_i$ for each $i = 1, \ldots, \ell$ is continuously differentiable, and $\nabla h_i(\bar{x})$ for $i = 1, \ldots, \ell$ are linearly independent. From the proof of Theorem 4.3.1 it follows that $G_0 \cap H_0 \subseteq A$. Thus, $\text{cl} (G_0 \cap H_0) \subseteq \text{cl} A \subseteq T \subseteq G' \cap H_0$. In particular, Cottle's constraint qualification implies that of Kuhn and Tucker, which in turn implies Abadie's constraint qualification.

We now demonstrate that Slater's constraint qualification and the linear independence constraint qualification imply that of Cottle. Suppose that Slater's qualification is satisfied, so that $g_i(x) < 0$ for $i \in I$ and $h_i(x) = 0$ for $i = 1, \ldots, \ell$ for some $x \in X$. By the pseudoconvexity of $g_i$ at $\bar{x}$, we get that $\nabla g_i(\bar{x})^\top (x - \bar{x}) < 0$ for $i \in I$.

Also, since $h_i(x) = h_i(\bar{x}) = 0$, the quasiconvexity and quasiconcavity of $h_i$ at $\bar{x}$ imply that $\nabla h_i(\bar{x})^\top (x - \bar{x}) = 0$. Letting $d = x - \bar{x}$, it follows that $d \in G_0 \cap H_0$. Thus, $G_0 \cap H_0 \neq \emptyset$, and the reader can verify that $\text{cl}(G_0 \cap H_0) = G' \cap H_0$. Therefore, Cottle's constraint qualification holds true.

Finally, we show that the linear independence constraint qualification implies that of Cottle. By contradiction, suppose that $G_0 \cap H_0 = \emptyset$. Then, using
a separation theorem as in the proof of Theorem 4.3.2, it follows that there exists a nonzero vector \((u, v)\) such that \(\sum_{i \in I} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^\ell v_i \nabla h_i(\bar{x}) = 0\), where \(u_i \geq 0\) is the vector whose \(i\)th component is \(u_i\) for \(i \in I\). This contradicts the linear independence assumption. Thus, Cottle's constraint qualification holds true.

In Figure 5.3 we summarize the implications of the constraint qualifications discussed above (see also Figure 5.2). As mentioned earlier, these implications, together with Theorem 5.3.1, validate the KKT conditions.

**Second-Order Constraint Qualifications for Inequality- and Equality-Constrained Problems**

In Chapter 4 we developed second-order necessary KKT optimality conditions. In particular, we observed in Theorem 4.4.3 that if \(\bar{x}\) is a local minimum and if all problem-defining functions are twice differentiable, with the gradients \(\nabla g_i(\bar{x})\) for \(i \in I\) and \(\nabla h_i(\bar{x})\) for \(i = 1, \ldots, \ell\) of the binding constraints being linearly independent, then \(\bar{x}\) is a KKT point and, additionally, \(d^T \nabla^2 L(\bar{x}) d \geq 0\) must hold true for all \(d \in C\) as defined therein. Hence, the linear independence condition affords a second-order constraint qualification, which implies that in addition to \(\bar{x}\) being a KKT point, a second-order type of condition must also hold true.

Alternatively, we can stipulate the following second-order constraint qualification, which is in the spirit of Abadie's constraint qualification. Suppose that all problem-defining functions are twice differentiable and that \(\bar{x}\) is a local

---

**Figure 5.3 Relationships among constraint qualifications for problems having inequality and equality constraints.**
Chapter 5

minimum at which Abadie's constraint qualification \( T = G' \cap H_0 \) holds true. Hence, we know from Theorem 5.3.1 that \( \bar{x} \) is a KKT point. Denote by \( \bar{u} \) and \( \bar{v} \) the associated set of Lagrangian multipliers corresponding to the inequality and equality constraints, respectively, and let \( I = \{ i : g_i(\bar{x}) = 0 \} \) represent the binding inequality constraints. Now, as in Theorem 4.4.3, define \( C = \{ d \neq 0 : \nabla g_i(\bar{x})'d = 0 \text{ for } i \in I^+, \nabla g_i(\bar{x})'d \leq 0 \text{ for } i \in I^0, \text{ and } \nabla h_i(\bar{x})'d = 0 \text{ for } i = 1,..., \ell \} \), where \( I^+ = \{ i \in I : \bar{u}_i > 0 \} \) and \( I^0 = I - I^+ \).

Accordingly, let \( T' \) denote the cone of tangents at \( \bar{x} \) when the inequality constraints with indices \( i \in I^+ \) are also treated as equalities. Then the stated second-order constraint qualification asserts that if \( T' = C \cup \{ 0 \} \), we must have \( d'\nabla^2 L(\bar{x})d \geq 0 \) for each \( d \in C \). We ask the reader to show that this assertion is valid in Exercise 5.9, using a proof similar to that of Theorem 4.4.3. Note that, in general, \( T' \subseteq C \cup \{ 0 \} \), in the same manner as \( T \subseteq G' \cap H_0 \). However, as evident from the proof of Theorem 4.4.3, under the linear independence constraint qualification, any \( d \in C \) also corresponds to a limiting direction based on a feasible arc and hence based on a sequence of points. Therefore, \( C \cup \{ 0 \} \subseteq T' \). This shows that the linear independence constraint qualification implies that \( T' = C \cup \{ 0 \} \). We ask the reader to construct the precise details of this argument in Exercise 5.9. In a similar manner, we can state another second-order constraint qualification in the spirit of Kuhn–Tucker's (cone of attainable directions) constraint qualification. This is addressed in Exercise 5.10, where we ask the reader to justify it and to show that this is also implied by the linear independence constraint qualification.

Exercises

[5.1] Prove that the cone of tangents defined in Definition 5.1.1 can be characterized equivalently in either of the following ways:

a. \( T = \{ d : \text{there exists a sequence } \{ \lambda_k \} \to 0^+ \text{ and a function } \alpha : \mathbb{R} \to \mathbb{R}^n, \text{ where } \alpha(\lambda) \to 0 \text{ as } \lambda \to 0, \text{ such that } x_k = \bar{x} + \lambda_k d + \lambda_k \alpha(\lambda_k) \in S \text{ for each } k \} \).

b. \( T = \{ d : d = \lambda \lim_{k \to \infty} \frac{x_k - \bar{x}}{\|x_k - \bar{x}\|}, \text{ where } \lambda \geq 0, \{ x_k \} \to \bar{x}, \text{ and where } x_k \in S \text{ and } x_k \neq \bar{x}, \text{ for each } k \} \).

[5.2] Prove that the cone of tangents is a closed cone. [Hint: First show that \( T = \bigcap_{N \in \mathcal{H}} \text{cl} K(S \cap N, \bar{x}) \), where \( K(S \cap N, \bar{x}) = \{ \lambda (x - \bar{x}) : x \in S \cap N, \lambda > 0 \} \) and \( \mathcal{H} \) is the class of all open neighborhoods about \( \bar{x} \).]

[5.3] For a nonlinear optimization problem, let \( \bar{x} \) be a feasible solution, let \( F_0 = \{ d : \nabla f(\bar{x})'d < 0 \} \), and let \( T \) be the cone
of tangents at \( \bar{x} \). If \( \bar{x} \) is a local minimum, is \( F \cap T = \emptyset \)? Is \( F \cap T = \emptyset \) sufficient to claim that \( \bar{x} \) is a local minimum? Give examples to justify your answers. Show that if there exists an \( \varepsilon \)-neighborhood about \( \bar{x} \) over which \( f \) is pseudoconvex and the feasible region is a convex set, then \( F_0 \cap T = \emptyset \) implies that \( \bar{x} \) is a local minimum, so these conditions also guarantee that \( \bar{x} \) is a local minimum whenever \( F \cap T = \emptyset \).

[5.4] Let \( S = \{ x \in X : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m \} \). Let \( \bar{x} \in S \), and let \( I = \{ i : g_i(x) = 0 \} \). Show that \( T \subseteq G' \), where \( T \) is the cone of tangents of \( S \) at \( \bar{x} \), and \( G' = \{ d : \nabla g_i(\bar{x})'d \leq 0 \text{ for } i \in I \} \).

[5.5] Consider the problem to maximize \( 5x - x^2 \) subject to \( g_1(x) \leq 0 \), where \( g_1(x) = x \).

a. Verify graphically that \( \bar{x} = 0 \) is the optimal solution.

b. Verify that each of the constraint qualifications discussed in Section 5.2 holds true at \( \bar{x} = 0 \).

c. Verify that the KKT necessary conditions hold true at \( \bar{x} = 0 \).

Now suppose that the constraint \( g_2(x) \leq 0 \) is added to the above problem, where

\[
g_2(x) = \begin{cases} 1 - x & \text{if } x \geq 0 \\ -1 - x & \text{if } x < 0. \end{cases}
\]

Note that \( \bar{x} = 0 \) is still the optimal solution and that \( g_2 \) is discontinuous and nonbinding at \( \bar{x} \). Check whether the constraint qualifications discussed in Section 5.2 and the KKT conditions hold true at \( \bar{x} = 0 \). (This exercise illustrates the need of the continuity assumption of the nonbinding constraints.)

[5.6] Let \( A \) be an \( m \times n \) matrix, and consider the cones \( G_0 = \{ d : Ad < 0 \} \) and \( G' = \{ d : Ad \leq 0 \} \). Prove that:

a. \( G_0 \) is an open convex cone.

b. \( G' \) is a closed convex cone.

c. \( G_0 = \text{int } G' \).

d. \( \text{cl } G_0 = G' \) if and only if \( G_0 \neq \emptyset \).

[5.7] Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \), \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \), and \( x \in X \), where \( X \) is open and where all problem-defining functions are differentiable. Let \( \bar{x} \) be a feasible solution. The Mangasarian–Fromovitz constraint qualification requires that \( \nabla h_i(\bar{x}) \) for \( i = 1, \ldots, \ell \) are linearly independent and that \( G_0 \cap H_0 \neq \emptyset \), where \( G_0 = \{ d : \nabla g_i(\bar{x})'d < 0 \text{ for } i \in I \} \), \( I = \{ i : g_i(\bar{x}) = 0 \} \), and \( H_0 = \{ d : \nabla h_i(\bar{x})'d = 0 \text{ for } i = 1, \ldots, \ell \} \). Show that \( G_0 \cap H_0 \neq \emptyset \) if and only if \( \text{cl} \{ G_0 \cap H_0 \} = G' \cap H_0 \), where \( G' = \{ d : \)}
\[ \nabla g_i(\bar{x})^t d \leq 0 \text{ for } i \in I \] and hence that this constraint qualification is equivalent to that of Cottle's.

[5.8] Let \( S \) be a subset of \( R^n \), and let \( \bar{x} \in \text{int } S \). Show that the cone of tangents of \( S \) at \( \bar{x} \) is \( R^n \).

[5.9] Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, ..., m \) and \( h_i(x) = 0 \) for \( i = 1, ..., \ell \), where all problem-defining functions are twice differentiable. Let \( \bar{x} \) be a local minimum, and suppose that the cone of tangents \( T = G' \cap H_0 \), where \( G' = \{ d : \nabla g_i(\bar{x})^t d \leq 0 \text{ for } i \in I \} \), \( I = \{ i : g_i(\bar{x}) = 0 \} \), and \( H_0 = \{ d : \nabla h_i(\bar{x})^t d = 0 \text{ for } i = 1, ..., \ell \} \). Hence, by Theorem 5.3.1, \( \bar{x} \) is a KKT point. Let \( \bar{u}_i, i = 1, ..., m \), and \( \bar{v}_i, i = 1, ..., \ell \), be the associated Lagrangian multipliers in the KKT solution with respect to the inequality and the equality constraints, and define \( I^+ = \{ i : \bar{u}_i > 0 \} \), and let \( I^0 = I - I^+ \). Now define \( T' \) as the cone of tangents at \( \bar{x} \) with respect to the region \( \{ x : g_i(x) = 0 \text{ for } i \in I^+ \}, \) \( g_i(x) \leq 0 \) for \( i \in I^0 \), \( h_i(x) = 0 \) for \( i = 1, ..., \ell \), and denote \( C = \{ d \neq 0 : \nabla g_i(\bar{x})^t d = 0 \text{ for } i \in I^+, \nabla g_i(\bar{x})^t d \leq 0 \text{ for } i \in I^0, \text{ and } \nabla h_i(\bar{x})^t d = 0 \text{ for } i = 1, ..., \ell \} \). Show that if \( T' = C \cup \{ 0 \} \), the second-order necessary condition \( d^t \nabla^2 L(\bar{x}) d \geq 0 \) holds true for all \( d \in C \), where \( L(x) = f(x) + \sum_{i \in I} \bar{u}_i g_i(x) + \sum_{i = 1}^{\ell} \bar{v}_i h_i(x) \). Also, show that the linear independence constraint qualification implies that \( T' = C \cup \{ 0 \} \). (Hint: Examine the proof of Theorem 4.4.3.)

[5.10] Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, ..., m \) and \( h_i(x) = 0 \) for \( i = 1, ..., \ell \), where all problem-defining functions are twice differentiable. Let \( \bar{x} \) be a local minimum that is also a KKT point. Define \( \bar{C} = \{ d \neq 0 : \nabla g_i(\bar{x})^t d = 0 \text{ for } i \in I, \nabla h_i(\bar{x})^t d = 0 \text{ for } i = 1, ..., \ell \} \), where \( I = \{ i : g_i(\bar{x}) = 0 \} \). The second-order cone of attainable directions constraint qualification is said to hold true at \( \bar{x} \) if every \( d \in \bar{C} \) is tangential to a twice differentiable arc incident at \( \bar{x} \); that is, for every \( d \in \bar{C} \), there exists a twice differentiable function \( \alpha : [0, \varepsilon] \to R^n \) for some \( \varepsilon > 0 \), such that for each \( 0 \leq \lambda \leq \varepsilon, \alpha(0) = \bar{x}, g_i(\alpha(\lambda)) = 0 \text{ for } i \in I, h_i(\alpha(\lambda)) = 0 \text{ for } i = 1, ..., \ell \), and \( \lim_{\lambda \to 0^+} [\alpha(\lambda) - \alpha(0)]/\lambda = \theta d \) for some \( \theta > 0 \). Assuming that this condition holds true, show that \( d^t \nabla^2 L(\bar{x}) d \geq 0 \) for all \( d \in \bar{C} \), where \( L(\bar{x}) \) is defined by (4.25). Also, show that this second-order constraint qualification is implied by the linear independence constraint qualification.
[5.11] Find the cone of tangents for each of the following sets at the point $\bar{x} = (0, 0)^t$:

a. $S = \{(x_1, x_2) : x_2 \geq -x_1^3\}$.

b. $S = \{(x_1, x_2) : x_1 \text{ is integer, } x_2 = 0\}$.

c. $S = \{(x_1, x_2) : x_1 \text{ is rational, } x_2 = 0\}$.

[5.12] Consider the problem to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$. Let $\bar{x}$ be feasible, and let $I = \{i : g_i(\bar{x}) = 0\}$. Let $(\bar{z}, \bar{d})$ be an optimal solution to the following linear program:

Minimize $z$

subject to $\nabla f(\bar{x})^t d - z \leq 0$

$\nabla g_i(\bar{x})^t d - z \leq 0$ for $i \notin I$

$-1 \leq d_j \leq 1$ for $j = 1, \ldots, n$.

a. Show that the Fritz John conditions hold true if $\bar{z} = 0$.

b. Show that if $\bar{z} = 0$, the KKT conditions hold true under Cottle's constraint qualification.

[5.13] Consider the following problem:

Minimize $-x_1$

subject to $x_1^2 + x_2^2 \leq 1$

$(x_1 - 1)^3 - x_2 \leq 0$.

a. Show that the Kuhn–Tucker constraint qualification holds true at $\bar{x} = (1, 0)^t$.

b. Show that $\bar{x} = (1, 0)^t$ is a KKT point and that it is the global optimal solution.

[5.14] For each of the following sets, find the cone of feasible directions and the cone of attainable directions at $\bar{x} = (0, 0)^t$:

a. $S = \{(x_1, x_2) : -1 \leq x_1 \leq 1, \quad x_2 \geq x_1^{1/3}, \quad x_2 \geq x_1\}$.

b. $S = \{(x_1, x_2) : x_2 > x_1^2\}$.

c. $S = \{(x_1, x_2) : x_2 = -x_1^3\}$.

d. $S = S_1 \cup S_2$, where

$S_1 = \{(x_1, x_2) : x_1 \geq 0, \quad x_2 \geq x_1^2\}$ and

$S_2 = \{(x_1, x_2) : x_1 \leq 0, \quad -2x_1 \leq 3x_2 \leq -x_1\}.$
[5.15] Consider the problem to minimize $f(x)$ subject to $x \in X$ and $g_i(x) \leq 0$ for $i = 1, \ldots, m$. Let $\overline{x}$ be a feasible point, and let $I = \{i : g_i(\overline{x}) = 0\}$. Suppose that $X$ is open and each $g_i$ for $i \notin I$ is continuous at $\overline{x}$. Further, suppose that the set

$$\{d : \nabla g_i(\overline{x})^t d \leq 0 \text{ for } i \in J, \nabla g_i(\overline{x})^t d < 0 \text{ for } i \in I - J\}$$

is not empty, where $J = \{i \in I : g_i$ is pseudoconcave at $\overline{x}\}$. Show that this condition is sufficient to validate the KKT conditions at $\overline{x}$. (This is the Arrow-Hurwicz-Uzawa constraint qualification.)

[5.16] Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at $\overline{x}$ with a nonzero gradient $\nabla f(\overline{x})$. Let $S = \{x : f(x) \geq f(\overline{x})\}$. Show that the cone of tangents and the cone of attainable directions of $S$ at $\overline{x}$ are both given by $\{d : \nabla f(\overline{x})^t d \geq 0\}$. Does this result hold true if $\nabla f(\overline{x}) = 0$? Prove or give a counterexample.

[5.17] Consider the feasible region $S = \{x \in X : g_1(x) \leq 0\}$, where $g_1(x) = x_1^2 + x_2^2 - 1$ and $X$ is the collection of all convex combinations of the four points $(-1,0)^t$, $(0,1)^t$, $(1,0)^t$, and $(0,-1)^t$.

a. Find the cone of tangents $T$ of $S$ at $\overline{x} = (1,0)^t$.

b. Check whether $T \supseteq G'$, where $G' = \{d : \nabla g_1(\overline{x})^t d \leq 0\}$.

c. Replace the set $X$ by four inequality constraints. Repeat parts a and b, where $G' = \{d : \nabla g_i(\overline{x})^t d \leq 0 \text{ for } i \in I\}$ and $I$ is the new set of binding constraints at $\overline{x} = (1,0)^t$.

[5.18] Let $S = \{x \in S : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m \text{ and } h_i(x) = 0 \text{ for } i = 1, \ldots, \ell\}$. Let $\overline{x} \in S$, and let $I = \{i : g_i(\overline{x}) = 0\}$. Show that $T \subseteq G' \cap H_0$, where $T$ is the cone of tangents of $S$ at $\overline{x}$, $G' = \{d : \nabla g_i(\overline{x})^t d \leq 0 \text{ for } i \in I\}$, and $H_0 = \{d : \nabla h_i(\overline{x})^t d = 0 \text{ for } i = 1, \ldots, \ell\}$.

[5.19] Consider Abadie's constraint qualification $T = G' \cap H_0$ for the case of inequality and equality constraints. Using the Kuhn-Tucker example of Figure 4.7, for instance, demonstrate by considering differentiable objective functions that it is valid and more general to instead require $F_0 \cap T = F_0 \cap G' \cap H_0$ to guarantee that if $\overline{x}$ is a local minimum, it is a KKT point. (Typically, "constraint qualifications" address only the behavior of the constraints and neglect the objective function.) Investigate the KKT conditions and Abadie's constraint qualification for the problem to minimize $\{f(x) : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m \text{ and } h_i(x) = 0 \text{ for } i = 1, \ldots, \ell\}$ versus those for the equivalent problem to minimize $\{z : f(x) \leq z, g_i(x) \leq 0 \text{ for } i = 1, \ldots, m \text{ and } h_i(x) = 0 \text{ for } i = 1, \ldots, \ell\}$. 

[5.20] Consider the constraints $Cd \leq 0$ and $d'd \leq 1$. Let $\bar{d}$ be a feasible solution such that $\bar{d}'\bar{d} = 1$, $C_1\bar{d} = 0$, and $C_2\bar{d} < 0$, where $C' = (C_1, C_2)$. Show that the constraint qualification $T = G_1 = \{d : C_1d \leq 0, d'd \leq 0\}$ holds true, where $T$ is the cone of tangents of the constraint set at $\bar{d}$.

[5.21] Consider the problem to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$ and $h_i(x) = 0$ for $i = 1, \ldots, l$, where all problem-defining functions are twice differentiable. Let $\bar{x}$ be a local minimum that also happens to be a KKT point with associated Lagrangian multipliers vectors $\bar{\mu}$ and $\bar{\nu}$ corresponding to the inequality and equality constraints, respectively. Define $I = \{i : g_i(\bar{x}) = 0\}$, $I^+ = \{i : \bar{\mu}_i > 0\}$, $I^0 = I - I^+$, $G_0 = \{d : \nabla g_i(\bar{x})'d < 0$ for $i \in I^0, \nabla g_i(\bar{x})'d = 0$ for $i \in I^+ \}$, and $H_0 = \{d : \nabla h_i(\bar{x})'d = 0$ for $i = 1, \ldots, l\}$. The strict Mangasarian-Fromovitz constraint qualification (SMFCQ) is said to hold true at $\bar{x}$ if $\nabla g_i(\bar{x})$ for $i \in I^+$ and $\nabla h_i(\bar{x})$ for $i = 1, \ldots, l$ are linearly independent and $G_0 \cap H_0 \neq \emptyset$. Show that the SMFCQ condition holds true at $\bar{x}$ if and only if the KKT Lagrangian multiplier vector $(\bar{\mu}, \bar{\nu})$ is unique. Moreover, show that if the SMFCQ condition holds true at $\bar{x}$, then $d'\nabla^2 L(\bar{x})d \geq 0$ for all $d \in C = \{d \neq 0 : \nabla g_i(\bar{x})'d = 0$ for $i \in I^+, \nabla g_i(\bar{x})'d \leq 0$ for $i \in I^0, \nabla h_i(\bar{x})'d = 0$ for $i = 1, \ldots, l\}$, where $L(x) = f(x) + \sum_{i \in I} \bar{\mu}_i g_i(x) + \sum_{i=1}^l \bar{\nu}_i h_i(x)$. (This result is due to Kyparisis [1985] and Ben-Tal [1980].)

[5.22] Consider the feasible region defined by $g_i(x) \leq 0$ for $i = 1, \ldots, m$, where $g_i : R^n \rightarrow R$, $i = 1, \ldots, m$, are differentiable functions. Let $\bar{x}$ be a feasible solution, denote by $\Xi$ the set of differentiable objective functions $f : R^n \rightarrow R$ for which $\bar{x}$ is a local minimum, and let $D\Xi = \{y : y = \nabla f(\bar{x})$ for some $f \in \Xi\}$. Define the set $G' = \{d : \nabla g_i(\bar{x})'d \leq 0$ for $i \in I\}$, where $I = \{i : g_i(\bar{x}) = 0\}$, and let $T$ be the cone of tangents at $\bar{x}$. Furthermore, for any set $S$, let $S_*$ denote its reverse polar cone defined as $\{y : y'x \geq 0$ for all $x \in S\}$.

a. Show that $D\Xi = T_*$.

b. Show that the KKT conditions hold true for all $f \in \Xi$ if and only if $T_* = G'_*$. (Hint: The statement in Part b occurs if and only if $D\Xi \subseteq G'_*$. Now, use Part a along with the fact that $T_* \supseteq G'_*$ since $T \subseteq G'$. This result is due to Gould and Tolle [1971, 1972].)
Notes and References

In this chapter we provide an alternative derivation of the KKT conditions for problems having inequality constraints and problems having both equality and inequality constraints. This is done directly by imposing a suitable constraint qualification, as opposed to first developing the Fritz John conditions and then the KKT conditions.

The KKT optimality conditions were originally developed by imposing the constraint qualification that for every direction vector $d$ in the cone $G'$, there is a feasible arc whose tangent at $\bar{x}$ points along $d$. Since then, many authors have developed the KKT conditions under different constraint qualifications. For a thorough study of this subject, refer to the works of Abadie [1967b], Arrow et al. [1961], Canon et al. [1966], Cottle [1963a], Evans [1970], Evans and Gould [1970], Guignard [1969], Mangasarian [1969a], Mangasarian and Fromovitz [1967], and Zangwill [1969]. For a comparison and further study of these constraint qualifications, see the survey articles of Bazaraa et al. [1972], Gould and Tolle [1972], and Peterson [1973].

Gould and Tolle [1971] showed that the constraint qualification of Guignard [1969] is the weakest possible in the sense that it is both necessary and sufficient for the validation of the KKT conditions (see Exercise 5.22 for a precise statement). For further discussion on constraint qualifications that validate second-order necessary optimality conditions, refer to Ben-Tal [1980], Ben-Tal and Zowe [1982], Fletcher [1987], Kyparisis [1985], and McCormick [1967]. Also, for an application of KKT conditions under various constraint qualifications in conducting sensitivity analyses in nonlinear programming, see Fiacco [1983].
Given a nonlinear programming problem, there is another nonlinear programming problem closely associated with it. The former is called the \textit{primal} problem, and the latter is called the \textit{Lagrangian dual problem}. Under certain convexity assumptions and suitable constraint qualifications, the primal and dual problems have equal optimal objective values and, hence, it is possible to solve the primal problem indirectly by solving the dual problem.

Several properties of the dual problem are developed in this chapter. They are used to provide general solution strategies for solving the primal and dual problems. As a by-product of one of the duality theorems, we obtain saddle point necessary optimality conditions without any differentiability assumptions.

Following is an outline of the chapter.

\textbf{Section 6.1: Lagrangian Dual Problem} \quad We introduce the Lagrangian dual problem, give its geometric interpretation, and illustrate it by several numerical examples.

\textbf{Section 6.2: Duality Theorems and Saddle Point Optimality Conditions} \quad We prove the weak and strong duality theorems. The latter shows that the primal and dual objective values are equal under suitable convexity assumptions. We also develop the saddle point optimality conditions along with necessary and sufficient conditions for the absence of a duality gap, and interpret this in terms of a suitable perturbation function.

\textbf{Section 6.3: Properties of the Dual Function} \quad We study several important properties of the dual function, such as concavity, differentiability, and subdifferentiability. We then give necessary and sufficient characterizations of ascent and steepest ascent directions.

\textbf{Section 6.4: Formulating and Solving the Dual Problem} \quad Several procedures for solving the dual problem are discussed. In particular, we describe briefly gradient and subgradient-based methods, and present a tangential approximation cutting plane algorithm.

\textbf{Section 6.5: Getting the Primal Solution} \quad We show that the points generated during the course of solving the dual problem yield optimal solutions to perturbations of the primal problem. For convex programs, we show how to obtain primal feasible solutions that are near-optimal.
Chapter 6

Section 6.6: Linear and Quadratic Programs

We give Lagrangian dual formulations for linear and quadratic programming, relating them to other standard duality formulations.

6.1 Lagrangian Dual Problem

Consider the following nonlinear programming Problem P, which we call the primal problem.

Primal Problem P

Minimize $f(x)$

subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$

$h_i(x) = 0$ for $i = 1, \ldots, \ell$

$x \in X$.

Several problems, closely related to the above primal problem, have been proposed in the literature and are called dual problems. Among the various duality formulations, the Lagrangian duality formulation has perhaps attracted the most attention. It has led to several algorithms for solving large-scale linear problems as well as convex and nonconvex nonlinear problems. It has also proved useful in discrete optimization where all or some of the variables are further restricted to be integers. The Lagrangian dual problem $D$ is stated below.

Lagrangian Dual Problem D

Maximize $\theta(u, v)$

subject to $u \geq 0$,

where $\theta(u, v) = \inf \{ f(x) + \sum_{i=1}^{m} u_i g_i(x) + \sum_{i=1}^{\ell} v_i h_i(x) : x \in X \}$.

Note that the Lagrangian dual function $\theta$ may assume the value of $-\infty$ for some vectors $(u, v)$. The optimization problem that evaluates $\theta(u, v)$ is sometimes referred to as the Lagrangian dual subproblem. In this problem the constraints $g_i(x) \leq 0$ and $h_i(x) = 0$ have been incorporated in the objective function using the Lagrangian multipliers or dual variables $u_i$ and $v_i$, respectively. This process of accommodating the constraints within the objective function using dual or Lagrangian multipliers is referred to as dualization. Also note that the multiplier $u_i$ associated with the inequality constraint $g_i(x) \leq 0$ is nonnegative, whereas the multiplier $v_i$ associated with the equality constraint $h_i(x) = 0$ is unrestricted in sign.

Since the dual problem consists of maximizing the infimum (greatest lower bound) of the function $f(x) + \sum_{i=1}^{m} u_i g_i(x) + \sum_{i=1}^{\ell} v_i h_i(x)$, it is sometimes referred to as the max–min dual problem. We remark here that strictly speaking,
we should write $D$ as $\sup \{ \theta(u, v) : u \geq 0 \}$, rather than $\max \{ \theta(u, v) : u \geq 0 \}$, since the maximum may not exist (see Example 6.2.8). However, we shall specifically identify such cases wherever necessary.

The primal and Lagrangian dual problems can be written in the following form using vector notation, where $f: \mathbb{R}^n \to \mathbb{R}$, $g: \mathbb{R}^n \to \mathbb{R}^m$ is a vector function whose $i$th component is $g_i$, and $h: \mathbb{R}^n \to \mathbb{R}^\ell$ is a vector function whose $i$th component is $h_i$. For the sake of convenience, we shall use this form throughout the remainder of this chapter.

**Primal Problem $P$**

Minimize $f(x)$
subject to $g(x) \leq 0$
$h(x) = 0$
$x \in X$.

**Lagrangian Dual Problem $D$**

Maximize $\theta(u, v)$
subject to $u \geq 0$, where $\theta(u, v) = \inf \{ f(x) + u^t g(x) + v^t h(x) : x \in X \}$.

Given a nonlinear programming problem, several Lagrangian dual problems can be devised, depending on which constraints are handled as $g(x) \leq 0$ and $h(x) = 0$ and which constraints are treated by the set $X$. This choice can affect both the optimal value of $D$ (as in nonconvex situations) and the effort expended in evaluating and updating the dual function $\theta$ during the course of solving the dual problem. Hence, an appropriate selection of the set $X$ must be made, depending on the structure of the problem and the purpose for solving $D$ (see the Notes and References section).

**Geometric Interpretation of the Dual Problem**

We now discuss briefly the geometric interpretation of the dual problem. For the sake of simplicity, we consider only one inequality constraint and assume that no equality constraints exist. Then the primal problem is to minimize $f(x)$ subject to $x \in X$ and $g(x) \leq 0$.

In the $(y, z)$ plane, the set $\{(y, z) : y = g(x), \ z = f(x) \text{ for some } x \in X\}$ is denoted by $G$ in Figure 6.1. Thus, $G$ is the image of $X$ under the $(g, f)$ map. The primal problem asks us to find a point in $G$ with $y \leq 0$ that has a minimum ordinate. Obviously, this point is $(\bar{y}, \bar{z})$ in Figure 6.1.
Now suppose that \( u \geq 0 \) is given. To determine \( \theta(u) \), we need to minimize \( f(x) + ug(x) \) over all \( x \in X \). Letting \( y = g(x) \) and \( z = f(x) \) for \( x \in X \), we want to minimize \( z + uy \) over points in \( G \). Note that \( z + uy = \alpha \) is an equation of a straight line with slope \(-u\) and intercept \( \alpha \) on the z-axis. To minimize \( z + uy \) over \( G \), we need to move the line \( z + uy = \alpha \) parallel to itself as far down (along its negative gradient) as possible while it remains in contact with \( G \). In other words, we move this line parallel to itself until it supports \( G \) from below, that is, the set \( G \) lies above the line and touches it. Then the intercept on the z-axis gives \( \theta(u) \), as shown in Figure 6.1. The dual problem is therefore equivalent to finding the slope of the supporting hyperplane such that its intercept on the z-axis is maximal. In Figure 6.1, such a hyperplane has slope \(-\bar{u}\) and supports the set \( G \) at the point \((\bar{y}, \bar{z})\). Thus, the optimal dual solution is \( \bar{u} \), and the optimal dual objective value is \( \bar{z} \). Furthermore, the optimal primal and dual objectives are equal in this case.

There is a related interesting interpretation that provides an important conceptual tool in this context. For the problem under consideration, define the function

\[
v(y) = \min \{ f(x) : g(x) \leq y, \ x \in X \}.
\]

The function \( v \) is called a perturbation function since it is the optimal value function of a problem obtained from the original problem by perturbing the right-hand side of the inequality constraint \( g(x) \leq 0 \) to \( y \) from the value of zero. Note that \( v(y) \) is a nonincreasing function of \( y \) since, as \( y \) increases, the feasible

---

Figure 6.1 Geometric interpretation of Lagrangian duality.
region of the perturbed problem enlarges (or stays the same). For the present case, this function is illustrated in Figure 6.1. Observe that \( v \) corresponds here to the lower envelope of \( G \) between points A and B because this envelope is itself monotone decreasing. Moreover, \( v \) remains constant at the value at point B for values of \( y \) higher than that at B and becomes \( \infty \) for points to the left of A because of infeasibility. In particular, if \( v \) is differentiable at the origin, we observe that \( v'(0) = -u \). Hence, the marginal rate of change in objective function value with an increase in the right-hand side of the constraint from its present value of zero is given by \( -u \), the negative of the Lagrangian multiplier value at optimality. If \( v \) is convex but is not differentiable at the origin, then \( -u \) is evidently a subgradient of \( v \) at \( y = 0 \). In either case we know that \( v(y) \geq v(0) - uy \) for all \( y \in \mathbb{R} \). As we shall see later, \( v \) can be nondifferentiable and/or nonconvex, but the condition \( v(y) \geq v(0) - uy \) holds true for all \( y \in \mathbb{R} \) if and only if \( u \) is a KKT Lagrangian multiplier corresponding to an optimal solution \( \mathbf{x} \) such that it solves the dual problem with equal primal and dual objective values. As seen above, this happens to be the case in Figure 6.1.

### 6.1.1 Example

Consider the following primal problem:

\[
\begin{align*}
\text{Minimize} & \quad x_1^2 + x_2^2 \\
\text{subject to} & \quad -x_1 - x_2 + 4 \leq 0 \\
& \quad x_1, x_2 \geq 0.
\end{align*}
\]

Note that the optimal solution occurs at the point \( (x_1, x_2) = (2, 2) \), whose objective value is equal to 8.

Letting \( g(x) = -x_1 - x_2 + 4 \) and \( X = \{(x_1, x_2) : x_1, x_2 \geq 0\} \), the dual function is given by

\[
\begin{align*}
\theta(u) &= \inf\{x_1^2 + x_2^2 + u(-x_1 - x_2 + 4) : x_1, x_2 \geq 0\} \\
&= \inf\{x_1^2 - ux_1 : x_1 \geq 0\} + \inf\{x_2^2 - ux_2 : x_2 \geq 0\} + 4u.
\end{align*}
\]

Note that the above infima are achieved at \( x_1 = x_2 = u/2 \) if \( u \geq 0 \) and at \( x_1 = x_2 = 0 \) if \( u < 0 \). Hence,

\[
\theta(u) = \begin{cases} 
-\frac{1}{2}u^2 + 4u & \text{for } u \geq 0 \\
4u & \text{for } u < 0.
\end{cases}
\]

Note that \( \theta \) is a concave function, and its maximum over \( u \geq 0 \) occurs at \( u = 4 \). Figure 6.2 illustrates the situation. Note also that the optimal primal and dual objective values are both equal to 8.
Now let us consider the problem in the \((y, z)-\)plane, where \(y = g(x)\) and \(z = f(x)\). We are interested in finding \(G\), the image of \(X = \{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0\}\), under the \((g, f)\) map. We do this by deriving explicit expressions for the lower and upper envelopes of \(G\), denoted, respectively, by \(\alpha\) and \(\beta\).

Given \(y\), note that \(\alpha(y)\) and \(\beta(y)\) are the optimal objective values of the following problems \(P_1\) and \(P_2\), respectively.

**Problem \(P_1\)**

Minimize \(x_1^2 + x_2^2\)  
subject to \(-x_1 - x_2 + 4 = y\)  
\(x_1, x_2 \geq 0\).

**Problem \(P_2\)**

Maximize \(x_1^2 + x_2^2\)  
subject to \(-x_1 - x_2 + 4 = y\)  
\(x_1, x_2 \geq 0\).
The reader can verify that $\alpha(y) = (4 - y)^2/2$ and $\beta(y) = (4 - y)^2$ for $y \leq 4$. The set $G$ is illustrated in Figure 6.2. Note that $x \in X$ implies that $x_1, x_2 \geq 0$, so that $-x_1 - x_2 + 4 \leq 4$. Thus, every point $x \in X$ corresponds to $y \leq 4$.

Note that the optimal dual solution is $\bar{u} = 4$, which is the negative of the slope of the supporting hyperplane shown in Figure 6.2. The optimal dual objective value is $\alpha(0) = 8$ and is equal to the optimal primal objective value.

Again, in Figure 6.2, the perturbation function $v(y)$ for $y \in R$ corresponds to the lower envelope $\alpha(y)$ for $y \leq 4$, and $v(y)$ remains constant at the value 0 for $y \geq 4$. The slope $v'(0)$ equals $-4$, the negative of the optimal Lagrange multiplier value. Moreover, we have $v(y) \geq v(0) - 4y$ for all $y \in R$. As we shall see in the next section, this is a necessary and sufficient condition for the primal and dual objective values to match at optimality.

### 6.2 Duality Theorems and Saddle Point Optimality Conditions

In this section we investigate the relationships between the primal and dual problems and develop saddle point optimality conditions for the primal problem.

Theorem 6.2.1, referred to as the weak duality theorem, shows that the objective value of any feasible solution to the dual problem yields a lower bound on the objective value of any feasible solution to the primal problem. Several important results follow as corollaries.

#### 6.2.1 Theorem (Weak Duality Theorem)

Let $x$ be a feasible solution to Problem $P$; that is, $x \in X$, $g(x) \leq 0$, and $h(x) = 0$. Also, let $(u, v)$ be a feasible solution to Problem $D$; that is, $u \geq 0$. Then $f(x) \geq \theta(u, v)$.

**Proof**

By the definition of $\theta$, and since $x \in X$, we have

$$\theta(u, v) = \inf \{f(y) + u^T g(y) + v^T h(y) : y \in X\}$$

$$\leq f(x) + u^T g(x) + v^T h(x) \leq f(x)$$

since $u \geq 0$, $g(x) \leq 0$, and $h(x) = 0$. This completes the proof.

**Corollary 1**

$$\inf \{f(x) : x \in X, \ g(x) \leq 0, \ h(x) = 0\} \geq \sup \{\theta(u, v) : u \geq 0\}.$$

**Corollary 2**

If $f(\bar{x}) = \theta(\bar{u}, \bar{v})$, where $\bar{u} \geq 0$ and $\bar{x} \in \{x \in X : g(x) \leq 0, \ h(x) = 0\}$, then $\bar{x}$ and $(\bar{u}, \bar{v})$ solve the primal and dual problems, respectively.
Corollary 3
If \(\inf\{f(x) : x \in X, g(x) \leq 0, h(x) = 0\} = -\infty\), then \(\theta(u, v) = -\infty\) for each \(u \geq 0\).

Corollary 4
If \(\sup\{\theta(u, v) : u \geq 0\} = \infty\), then the primal problem has no feasible solution.

Duality Gap
From Corollary 1 to Theorem 6.2.1, the optimal objective value of the primal problem is greater than or equal to the optimal objective value of the dual problem. If strict inequality holds true, a duality gap is said to exist. Figure 6.3 illustrates the case of a duality gap for a problem having a single inequality constraint and no equality constraints. The perturbation function \(v(y)\) for \(y \in \mathbb{R}\) is as shown in the figure. Note that by definition, this is the greatest monotone nonincreasing function that envelopes \(G\) from below (see Exercise 6.1). The optimal primal value is \(v(0)\). The greatest intercept on the ordinate \(z\)-axis achieved by a hyperplane that supports \(G\) from below gives the optimal dual objective value as shown. In particular, observe that there does not exist a \(u^*\) such that \(v(y) \geq v(0) - u^*y\) for all \(y \in \mathbb{R}\), as we had in Figures 6.1 and 6.2.

Exercise 6.2 asks the reader to construct \(G\) and \(v\) for the instance illustrated in Figure 4.13 that results in a situation similar to that of Figure 6.3.

6.2.2 Example
Consider the following problem:

Figure 6.3 Duality gap.
Minimize \( f(x) = -2x_1 + x_2 \)
subject to \( h(x) = x_1 + x_2 - 3 = 0 \)
\((x_1, x_2) \in X, \)

where \( X = \{(0, 0), (0, 4), (4, 4), (4, 0), (1, 2), (2, 1)\}. \)

It is easy to verify that \((2, 1)\) is the optimal solution to the primal problem with objective value equal to \(-3\). The dual objective function \( \theta \) is given by

\[
\theta(v) = \min \{( -2x_1 + x_2 ) + v(x_1 + x_2 - 3) : (x_1, x_2) \in X \}.
\]

The reader may verify that the explicit expression for \( \theta \) is given by

\[
\theta(v) = \begin{cases} 
14 + 5v & \text{for } v \leq -1 \\
-8 + v & \text{for } -1 \leq v \leq 2 \\
-3v & \text{for } v \geq 2.
\end{cases}
\]

The dual function is shown in Figure 6.4, and the optimal solution is \( \bar{v} = 2 \) with objective value \(-6\). Note that there exists a duality gap in this example.

In this case, the set \( G \) consists of a finite number of points, each corresponding to a point in \( X \). This is shown in Figure 6.5. The supporting hyperplane, whose intercept on the vertical axis is maximal, is shown in the figure. Note that the intercept is equal to \(-6\) and that the slope is equal to \(-2\). Thus, the optimal dual solution is \( \bar{v} = 2 \), with objective value \(-6\). Furthermore, note that the points in the set \( G \) on the vertical axis correspond to the primal feasible points and, hence, the minimal primal objective value is equal to \(-3\).

Similar to the inequality constrained case, the perturbation function here is defined as \( \nu(y) = \min \{ f(x) : h(x) = y, \ x \in X \} \). Because of the discrete nature of \( X \), \( h(x) \) can take on only a finite possible number of values. Hence,
noting $G$ in Figure 6.5, we obtain $v(-3) = 0$, $v(0) = -3$, $v(1) = -8$, and $v(5) = -4$, with $v(y) = \infty$ for all $y \in R$ otherwise. Again, the optimal primal value is $v(0) = -3$, and there does not exist a $\bar{y}$ such that $v(y) \geq v(0) - \bar{y}$. Hence, a duality gap exists.

Conditions that guarantee the absence of a duality gap are given in Theorem 6.2.4. Then Theorem 6.2.7 relates this to the perturbation function. First, however, the following lemma is needed.

6.2.3 Lemma

Let $X$ be a nonempty convex set in $R^n$. Let $\alpha: R^n \rightarrow R$ and $g: R^n \rightarrow R^m$ be convex, and let $h: R^n \rightarrow R^\ell$ be affine; that is, $h$ is of the form $h(x) = Ax - b$. If System 1 below has no solution $x$, then System 2 has a solution $(u_0, u, v)$. The converse holds true if $u_0 > 0$.

**System 1:** $\alpha(x) < 0$, $g(x) \leq 0$, $h(x) = 0$ for some $x \in X$

**System 2:** $u_0 \alpha(x) + u^T g(x) + v^T h(x) \geq 0$ for all $x \in X$

$(u_0, u) \geq 0$, $(u_0, u, v) \neq 0$.

**Proof**

Suppose that System 1 has no solution, and consider the following set:

$\Lambda = \{(p, q, r): p > \alpha(x), q \geq g(x), r = h(x) \text{ for some } x \in X\}$.

![Figure 6.5 Geometric interpretation of Example 6.2.2.](image-url)
Lagrangian Duality and Saddle Point Optimality Conditions

Noting that $X$, $\alpha$, and $g$ are convex and that $h$ is affine, it can easily be shown that $\Lambda$ is convex. Since System 1 has no solution, $(0, 0, 0) \notin \Lambda$. By Corollary 1 to Theorem 2.4.7, there exists a nonzero $(u_0, u, v)$ such that

$$u_0 p + u' q + v' r \geq 0 \quad \text{for each } (p, q, r) \in \text{cl } \Lambda. \quad (6.1)$$

Now fix an $x \in X$. Since $p$ and $q$ can be made arbitrarily large, (6.1) holds true only if $u_0 \geq 0$ and $u \geq 0$. Furthermore, $(p, q, r) = [\alpha(x), g(x), h(x)]$ belongs to $\text{cl } \Lambda$. Therefore, from (6.1), we get

$$u_0 \alpha(x) + u' g(x) + v' h(x) \geq 0.$$  

Since the above inequality is true for each $x \in X$, System 2 has a solution.

To prove the converse, assume that System 2 has a solution $(u_0, u, v)$ such that $u_0 > 0$ and $u \geq 0$, satisfying

$$u_0 \alpha(x) + u' g(x) + v' h(x) \geq 0 \quad \text{for each } x \in X.$$  

Now let $x \in X$ be such that $g(x) \leq 0$ and $h(x) = 0$. From the above inequality, since $u \geq 0$, we conclude that $u_0 \alpha(x) \geq 0$. Since $u_0 > 0$, $\alpha(x) > 0$; and hence, System 1 has no solution. This completes the proof.

Theorem 6.2.4, referred to as the strong duality theorem, shows that under suitable convexity assumptions and under a constraint qualification, the optimal objective function values of the primal and dual problems are equal.

6.2.4 Theorem (Strong Duality Theorem)

Let $X$ be a nonempty convex set in $R^n$, let $f: R^n \to R$ and $g: R^n \to R^m$ be convex, and let $h: R^n \to R^l$ be affine; that is, $h$ is of the form $h(x) = Ax - b$. Suppose that the following constraint qualification holds true. There exists an $\tilde{x} \in X$ such that $g(\tilde{x}) < 0$ and $h(\tilde{x}) = 0$, and $0 \in \text{int } h(X)$, where $h(X) = \{h(x) : x \in X\}$. Then

$$\inf\{f(x) : x \in X, g(x) \leq 0, h(x) = 0\} = \sup\{\theta(u, v) : u \geq 0\}. \quad (6.2)$$

Furthermore, if the inf is finite, then $\sup\{\theta(u, v) : u \geq 0\}$ is achieved at $(\bar{u}, \bar{v})$ with $\bar{u} \geq 0$. If the inf is achieved at $\tilde{x}$, then $\bar{u}' g(\tilde{x}) = 0$.

Proof

Let $\gamma = \inf\{f(x) : x \in X, g(x) \leq 0, h(x) = 0\}$. By assumption, $\gamma < \infty$. If $\gamma = -\infty$, then by Corollary 3 to Theorem 6.2.1, $\sup\{\theta(u, v) : u \geq 0\} = -\infty$, and therefore (6.2) holds true. Hence, suppose that $\gamma$ is finite, and consider the following system:
\[ f(x) - \gamma < 0, \quad g(x) \leq 0, \quad h(x) = 0, \quad x \in X. \]

By the definition of \( \gamma \), this system has no solution. Hence, from Lemma 6.2.3, there exists a nonzero vector \((u_0, u, v)\) with \((u_0, u, v) \geq 0\) such that

\[ u_0[f(x)-\gamma] + u'g(x) + v'h(x) \geq 0 \quad \text{for all } x \in X. \tag{6.3} \]

We first show that \( u_0 > 0 \). By contradiction, suppose that \( u_0 = 0 \). By assumption there exists an \( \hat{x} \in X \) such that \( g(\hat{x}) < 0 \) and \( h(\hat{x}) = 0 \). Substituting in (6.3), it follows that \( u'g(\hat{x}) \geq 0 \). Since \( g(\hat{x}) < 0 \) and \( u \geq 0 \), \( u'g(\hat{x}) \geq 0 \) is possible only if \( u = 0 \). But from (6.3), \( u_0 = 0 \) and \( u = 0 \), which implies that \( v'h(x) \geq 0 \) for all \( x \in X \). But since \( 0 \in \text{int} \ h(X) \), we can pick an \( x \in X \) such that \( h(x) = -\lambda v \), where \( \lambda > 0 \). Therefore, \( 0 \leq v'h(x) = -\lambda \|v\|^2 \), which implies that \( v = 0 \). Thus, we have shown that \( u_0 = 0 \) implies that \((u_0, u, v) = 0\), which is impossible. Hence, \( u_0 > 0 \). Dividing (6.3) by \( u_0 \) and denoting \( u/u_0 \) and \( v/u_0 \) by \( \bar{u} \) and \( \bar{v} \), respectively, we get

\[ f(x) + \bar{u}'g(x) + \bar{v}'h(x) \geq \gamma \quad \text{for all } x \in X. \tag{6.4} \]

This shows that \( \theta(\bar{u}, \bar{v}) = \inf \{f(x) + \bar{u}'g(x) + \bar{v}'h(x) : x \in X \} \geq \gamma \). In view of Theorem 6.2.1, it is then clear that \( \theta(\bar{u}, \bar{v}) = \gamma \), and \((\bar{u}, \bar{v})\) solves the dual problem.

To complete the proof, suppose that \( \bar{x} \) is an optimal solution to the primal problem; that is, \( \bar{x} \in X \), \( g(\bar{x}) \leq 0 \), \( h(\bar{x}) = 0 \), and \( f(\bar{x}) = \gamma \). From (6.4), letting \( x = \bar{x} \), we get \( \bar{u}'g(\bar{x}) \geq 0 \). Since \( \bar{u} \geq 0 \) and \( g(\bar{x}) \leq 0 \), we get \( \bar{u}'g(\bar{x}) = 0 \), and the proof is complete.

In Theorem 6.2.4, the assumption \( 0 \in \text{int} h(X) \) and that there exists an \( \hat{x} \in X \) such that \( g(\hat{x}) < 0 \) and \( h(\hat{x}) = 0 \) can be viewed as a generalization of Slater's constraint qualification of Chapter 5. In particular, if \( X = \mathbb{R}^n \), then \( 0 \in \text{int} h(X) \) holds true automatically (if redundant equations are deleted), so that the constraint qualification asserts the existence of a point \( \hat{x} \) such that \( g(\hat{x}) < 0 \) and \( h(\hat{x}) = 0 \). To see this, suppose that \( h(x) = Ax - b \). Without loss of generality, assume that rank \((A) = m \), because otherwise, any redundant constraints could be deleted. Now, any \( y \in \mathbb{R}^m \) could be represented as \( y = Ax - b \), where \( x = A'(AA')^{-1}(y + b) \). Thus, \( h(X) = \mathbb{R}^m \) and, in particular, \( 0 \in \text{int} h(X) \).
Saddle Point Criteria

The foregoing theorem shows that under convexity assumptions and under a suitable constraint qualification, the primal and dual objective function values match at optimality. Actually, a necessary and sufficient condition for the latter property to hold true is the existence of a saddle point, as we learn next. Given the primal Problem P, define the Lagrangian function

$$\phi(x, u, v) = f(x) + u'g(x) + v'h(x).$$

A solution \((x, \bar{u}, \bar{v})\) is called a saddle point of the Lagrangian function if \(x \in X\), \(\bar{u} \geq 0\), and

$$\phi(x, u, v) \leq \phi(x, \bar{u}, \bar{v}) \leq \phi(x, \bar{u}, v)$$

for all \(x \in X\), and all \((u, v)\) with \(u \geq 0\). (6.5)

Hence, we have that \(x\) minimizes \(\phi\) over \(X\) when \((u, v)\) is fixed at \((\bar{u}, \bar{v})\), and that \((\bar{u}, \bar{v})\) maximizes \(\phi\) over all \((u, v)\) with \(\bar{u} \geq 0\) when \(x\) is fixed at \(x\). Relating this to Figure 4.2, we see why \((\bar{x}, \bar{u}, \bar{v})\) is called a saddle point for the Lagrangian function \(\phi\).

The following result characterizes a saddle point solution and shows that its existence is a necessary and sufficient condition for the absence of a duality gap.

6.2.5 Theorem (Saddle Point Optimality and Absence of a Duality Gap)

A solution \((\bar{x}, \bar{u}, \bar{v})\) with \(\bar{x} \in X\) and \(\bar{u} \geq 0\) is a saddle point for the Lagrangian function \(\phi(x, u, v) = f(x) + u'g(x) + v'h(x)\) if and only if

a. \(\phi(\bar{x}, \bar{u}, \bar{v}) = \min\{\phi(x, \bar{u}, \bar{v}) : x \in X\}\),

b. \(g(\bar{x}) \leq 0\), \(h(\bar{x}) = 0\), and

c. \(\bar{u}'g(\bar{x}) = 0\).

Moreover, \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point if and only if \(\bar{x}\) and \((\bar{u}, \bar{v})\) are, respectively, optimal solutions to the primal and dual problems P and D with no duality gap, that is, with \(f(\bar{x}) = \theta(\bar{u}, \bar{v})\).

Proof

Suppose that \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point for the Lagrangian function \(\phi\). By definition, Condition (a) must be true. Furthermore, from (6.5), we have

$$f(\bar{x}) + \bar{u}'g(\bar{x}) + \bar{v}'h(\bar{x}) \geq f(x) + u'g(x) + v'h(x)$$

for all \((u, v)\) with \(u \geq 0\). (6.6)
Clearly, this implies that we must have \( g(\bar{x}) \leq 0 \) and \( h(\bar{x}) = 0 \), or else \((6.6)\) can be violated by appropriately making a component of \( u \) or \( v \) sufficiently large in magnitude. Now, taking \( u = 0 \) in \((6.6)\), we obtain that \( \bar{u}'g(\bar{x}) \geq 0 \). Noting that \( \bar{u} \geq 0 \) and \( g(\bar{x}) \leq 0 \) imply that \( \bar{u}'g(\bar{x}) \leq 0 \), we must have \( \bar{u}'g(\bar{x}) = 0 \). Hence, conditions \((a)\), \((b)\), and \((c)\) hold true.

Conversely, suppose that we are given \((\bar{x}, \bar{u}, \bar{v})\) with \( \bar{x} \in X \) and \( \bar{u} \geq 0 \) such that conditions \((a)\), \((b)\), and \((c)\) hold true. Then \( \phi(\bar{x}, \bar{u}, \bar{v}) = \phi(x, \bar{u}, \bar{v}) \) for all \( x \in X \) by Property \((a)\). Furthermore, \( \phi(\bar{x}, \bar{u}, \bar{v}) = f(\bar{x}) + \bar{u}'g(\bar{x}) + \bar{v}'h(\bar{x}) = f(\bar{x}) \geq f(x) + u'g(x) + v'h(x) = \phi(x, u, v) \) for all \((u, v)\) with \( u \geq 0 \), since \( g(\bar{x}) \leq 0 \) and \( h(\bar{x}) = 0 \). Hence, \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point. This proves the first part of the theorem.

Next, suppose again that \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point. By Property \((b)\), \( \bar{x} \) is feasible to Problem \( P \). Since \( \bar{u} \geq 0 \), we also have that \((\bar{u}, \bar{v})\) is feasible to \( D \). Moreover, by properties \((a)\), \((b)\), and \((c)\), \( \theta(\bar{u}, \bar{v}) = \phi(\bar{x}, \bar{u}, \bar{v}) = f(\bar{x}) + \bar{u}'g(\bar{x}) + \bar{v}'h(\bar{x}) = f(\bar{x}) \) by Corollary 2 to Theorem 6.2.1, \( \bar{x} \) and \((\bar{u}, \bar{v})\) solve \( P \) and \( D \), respectively, with no duality gap.

Finally, suppose that \( \bar{x} \) and \((\bar{u}, \bar{v})\) are optimal solutions to problems \( P \) and \( D \), respectively, with \( f(\bar{x}) = \theta(\bar{u}, \bar{v}) \). Hence, we have \( \bar{x} \in X \), \( g(\bar{x}) \leq 0 \), \( h(\bar{x}) = 0 \), and \( \bar{u} \geq 0 \). Moreover, we have by primal–dual feasibility that

\[
\theta(\bar{u}, \bar{v}) = \min \{ f(x) + \bar{u}'g(x) + \bar{v}'h(x) : x \in X \}
\leq f(\bar{x}) + \bar{u}'g(\bar{x}) + \bar{v}'h(\bar{x}) = f(\bar{x})
\]

But \( \theta(\bar{u}, \bar{v}) = f(\bar{x}) \), by hypothesis. Hence, equality holds true throughout the discussion above. In particular, \( \bar{u}'g(\bar{x}) = 0 \), so \( \phi(\bar{x}, \bar{u}, \bar{v}) = f(\bar{x}) = \theta(\bar{u}, \bar{v}) = \min \{ \phi(x, \bar{u}, \bar{v}) : x \in X \} \). Hence, properties \((a)\), \((b)\), and \((c)\) hold true in addition to \( \bar{x} \in X \) and \( \bar{u} \geq 0 \); so \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point. This completes the proof.

**Corollary**

Suppose that \( X, f, \) and \( g \) are convex and that \( h \) is affine; that is, \( h \) is of the form \( h(x) = Ax - b \). Further, suppose that \( 0 \in \text{int} h(X) \) and that there exists an \( \hat{x} \in X \) with \( g(\hat{x}) < 0 \) and \( h(\hat{x}) = 0 \). If \( \bar{x} \) is an optimal solution to the primal Problem \( P \), there exists a vector \((\bar{u}, \bar{v})\) with \( \bar{u} \geq 0 \) such that \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point.
Lagrangian Duality and Saddle Point Optimality Conditions

Proof

By Theorem 6.2.4 there exists an optimal solution \((\bar{u}, \bar{v})\), \(\bar{u} \geq 0\) to Problem D such that \(f(\bar{x}) = \theta(\bar{u}, \bar{v})\). Hence, by Theorem 6.2.5, \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point solution. This completes the proof.

There is an additional insight that can be derived in regard to the duality gap between the primal and dual problems. Note that the dual problem's optimal value is given by

\[
\theta^* = \sup_{(u, v): u \geq 0} \inf_{x \in X} [\phi(x, u, v)].
\]

If we interchange the order of optimization (see Exercise 6.3), we get

\[
\theta^* \leq \inf_{x \in X} \sup_{(u, v): u \geq 0} [\phi(x, u, v)].
\]

But the supremum of \(\phi(x, u, v) = f(x) + u^t g(x) + v^t h(x)\) over \((u, v)\) with \(u \geq 0\) is infinity, unless \(g(x) \leq 0\) and \(h(x) = 0\), whence it is \(f(x)\). Hence,

\[
\theta^* \leq \inf_{x \in X} \sup_{(u, v): u \geq 0} [\phi(x, u, v)]
\]

\[
= \inf \{f(x): g(x) \leq 0, h(x) = 0, x \in X\},
\]

which is the primal optimal value. Hence, we see that the primal and dual objective values match at optimality if and only if the interchange of the foregoing infimum and supremum operations leaves the optimal values unchanged. By Theorem 6.2.5, assuming that an optimum exists, this occurs if and only if there exists a saddle point \((\bar{x}, \bar{u}, \bar{v})\) for the Lagrangian function \(\phi\).

Relationship Between the Saddle Point Criteria and the Karush–Kuhn–Tucker Conditions

In Chapters 4 and 5, we discussed the KKT optimality conditions for Problem P:

Minimize \(f(x)\)
subject to \(g(x) \leq 0\)
\(h(x) = 0\)
\(x \in X\).

Furthermore, in Theorem 6.2.5 we developed the saddle point optimality conditions for the same problem. Theorem 6.2.6 gives the relationship between these two types of optimality conditions.
6.2.6 Theorem

Let \( S = \{x \in X : g(x) \leq 0, \ h(x) = 0\} \), and consider Problem P to minimize \( f(x) \) subject to \( x \in S \). Suppose that \( \bar{x} \in S \) satisfies the KKT conditions; that is, there exist \( \bar{u} \geq 0 \) and \( \bar{v} \) such that

\[
\nabla f(\bar{x}) + \nabla g(\bar{x})^T \bar{u} + \nabla h(\bar{x})^T \bar{v} = 0
\]
\[
\bar{u}^T g(\bar{x}) = 0.
\]

Suppose that \( f \) and \( g_i \) for \( i \in I \) are convex at \( \bar{x} \), where \( I = \{i : g_i(\bar{x}) = 0\} \). Further, suppose that if \( \nabla_i x \neq 0 \), then \( h_i \) is affine. Then \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point for the Lagrangian function \( \phi(x, u, v) = f(x) + u^T g(x) + v^T h(x) \).

Conversely, suppose that \((\bar{x}, \bar{u}, \bar{v})\) with \( \bar{x} \in \text{int} X \) and \( \bar{u} \geq 0 \) is a saddle point solution. Then \( \bar{x} \) is feasible to Problem P, and furthermore, \((\bar{x}, \bar{u}, \bar{v})\) satisfies the KKT conditions specified by (6.7).

Proof

Suppose that \((\bar{x}, \bar{u}, \bar{v})\), with \( \bar{x} \in S \) and \( \bar{u} \geq 0 \), satisfies the KKT conditions specified by (6.7). By convexity at \( \bar{x} \) of \( f \) and \( g_i \) for \( i \in I \), and since \( h_i \) is affine for \( \bar{x} \), we get

\[
f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x})
\]
\[
g_i(x) \geq g_i(\bar{x}) + \nabla g_i(\bar{x})^T (x - \bar{x}) \quad \text{for } i \in I
\]
\[
h_i(x) = h_i(\bar{x}) + \nabla h_i(\bar{x})^T (x - \bar{x}) \quad \text{for } i = 1, \ldots, \ell, \ \nabla_i x \neq 0
\]
for all \( x \in X \). Multiplying (6.8b) by \( \bar{u}_i \geq 0 \), (6.8c) by \( \nabla_i x \), and adding these to (6.8a) and noting (6.7), it follows from the definition of \( \phi \) that \( \phi(x, u, v) \geq \phi(\bar{x}, \bar{u}, \bar{v}) \) for all \( x \in X \). Also, since \( g(\bar{x}) \leq 0 \), \( h(\bar{x}) = 0 \), and \( \bar{u}^T g(\bar{x}) = 0 \), it follows that \( \phi(\bar{x}, u, v) \leq \phi(\bar{x}, \bar{u}, \bar{v}) \) for all \((u, v)\) with \( u \geq 0 \). Hence, \((\bar{x}, \bar{u}, \bar{v})\) satisfies the saddle point conditions given by (6.5).

To prove the converse, suppose that \((\bar{x}, \bar{u}, \bar{v})\) with \( \bar{x} \in \text{int} X \) and \( \bar{u} \geq 0 \) is a saddle point solution. Since \( \phi(\bar{x}, u, v) \leq \phi(\bar{x}, \bar{u}, \bar{v}) \) for all \( u \geq 0 \) and all \( v \), we have, using (6.6) as in Theorem 6.2.5, that \( g(\bar{x}) \leq 0 \), \( h(\bar{x}) = 0 \), and \( \bar{u}^T g(\bar{x}) = 0 \). This shows that \( \bar{x} \) is feasible to Problem P. Since \( \phi(\bar{x}, u, v) \leq \phi(x, u, v) \) for all \( x \in X \), then \( \bar{x} \) solves the problem to minimize \( \phi(x, u, v) \) subject to \( x \in X \). Since \( \bar{x} \in \text{int} X \), then \( \nabla_x \phi(\bar{x}, \bar{u}, \bar{v}) = 0 \), that is, \( \nabla f(\bar{x}) + \nabla g(\bar{x})^T \bar{u} + \nabla h(\bar{x})^T \bar{v} = 0 \); and hence, (6.7) holds true. This completes the proof.
Theorem 6.2.6 shows that if \( \bar{x} \) is a KKT point, then under certain convexity assumptions, the Lagrangian multipliers in the KKT conditions also serve as the multipliers in the saddle point criteria. Conversely, the multipliers in the saddle point conditions are the Lagrangian multipliers of the KKT conditions. Moreover, in view of Theorems 6.2.4, 6.2.5, and 6.2.6, the optimal dual variables for the Lagrangian dual problem are precisely the Lagrangian multipliers for the KKT conditions and also the multipliers for the saddle point conditions in this case.

**Saddle Point Optimality Interpretation Using a Perturbation Function**

While discussing the geometric interpretation of the dual problem and the associated duality gap, we introduced the concept of a perturbation function \( v \) and illustrated this in Examples 6.1.1 and 6.1.2 (see Figures 6.1 through 6.5). As alluded to previously, the existence of a supporting hyperplane to the epigraph of this function at the point \((0, v(0))\) related to the absence of a duality gap in these examples. This is formalized in the discussion that follows.

Consider the primal Problem P, and define the perturbation function \( v: R^{m+\ell} \to R \) as the optimal value function of the following problem, where \( y = (y_1, \ldots, y_m, y_{m+1}, \ldots, y_{m+\ell}) \):

\[
v(y) = \min \{ f(x) : g_i(x) \leq y_i \text{ for } i = 1, \ldots, m, \quad h_i(x) = y_{m+i} \text{ for } i = 1, \ldots, \ell, \ x \in X \}. \tag{6.9}
\]

Theorem 6.2.7 asserts that if Problem P has an optimum, the existence of a saddle point solution, that is, the absence of a duality gap, is equivalent to the existence of a supporting hyperplane for the epigraph of \( v \) at the point \((0, v(0))\).

**6.2.7 Theorem**

Consider the primal Problem P, and assume that an optimal solution \( \bar{x} \) to this problem exists. Then \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point for the Lagrangian function \( \phi(x, u, v) = f(x) + u^t g(x) + v^t h(x) \) if and only if

\[
v(y) \geq v(0) - (\bar{u}^t, \bar{v}^t)y \quad \text{for all } y \in R^{m+\ell}, \tag{6.10}
\]

that is, if and only if the hyperplane \( z = v(0) - (\bar{u}^t, \bar{v}^t)y \) supports the epigraph \( \{ (y, z) : z \geq v(y), y \in R^{m+\ell} \} \) of \( v \) at the point \((y, z) = (0, v(0))\).

**Proof**

Suppose that \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point solution. Then, by Theorem 6.2.5, the absence of a duality gap asserts that...
\[ v(0) = \theta(\bar{u}, \bar{v}) = \min \{ f(x) + \bar{u}'g(x) + \bar{V}'h(x) : x \in X \} \]

\[ = (\bar{u}', \bar{V}')y + \min \left\{ f(x) + \sum_{i=1}^{m} \bar{u}_i [g_i(x) - y_i] + \sum_{i=1}^{l} \bar{v}_i [h_i(x) - y_{m+i}] : x \in X \right\} \text{ for any } y \in \mathbb{R}^{m+l}. \]

Applying the weak duality Theorem 6.2.1 to the perturbed problem (6.9), we obtain from the foregoing identity that \( v(0) \leq (\bar{u}', \bar{V}')y + v(y) \) for any \( y \in \mathbb{R}^{m+l} \), so (6.10) holds true.

Conversely, suppose that (6.10) holds true for some \((\bar{u}, \bar{V})\), and let \( \bar{x} \) solve Problem P. We must show that \((\bar{x}, \bar{u}, \bar{V})\) is a saddle point solution. First, note that \( \bar{x} \in X \), \( g(\bar{x}) \leq 0 \), and \( h(\bar{x}) = 0 \). Furthermore, \( \bar{u} \geq 0 \) must hold true, because if \( \bar{u}_p < 0 \), say, then by selecting \( y \) such that \( y_i = 0 \) for \( i \neq p \), and \( y_p > 0 \), we obtain \( v(0) \geq v(y) \geq v(0) - \bar{u}_p y_p \), which implies that \( \bar{u}_p y_p \geq 0 \), a contradiction.

Second, observe that by fixing \( y = \bar{y} \equiv [g(\bar{x})', h(\bar{x})'] \) in (6.9), we obtain a restriction of Problem P, since \( g(\bar{x}) \leq 0 \) and \( h(\bar{x}) = 0 \). But for the same reason, since \( \bar{x} \) is feasible to (6.9) with \( y \) fixed as such, and since \( \bar{x} \) solves Problem P, we obtain \( v(\bar{y}) = v(0) \). By (6.10), this in turn means that \( \bar{u}'g(\bar{x}) \geq 0 \). Since \( g(\bar{x}) \leq 0 \) and \( \bar{u} \geq 0 \), we therefore must have \( \bar{u}'g(\bar{x}) = 0 \).

Finally, we have

\[ \phi(\bar{x}, \bar{u}, \bar{V}) = f(\bar{x}) + \bar{u}'g(\bar{x}) + \bar{V}'h(\bar{x}) = f(\bar{x}) = v(0) \leq v(y) + (\bar{u}', \bar{V}')y \] (6.11)

for all \( y \in \mathbb{R}^{m+l} \). Now, for any \( \hat{x} \in X \), denoting \( \hat{y} = [g(\hat{x})', h(\hat{x})'] \), we obtain from (6.9) that \( v(\hat{y}) \leq f(\hat{x}) \), since \( \hat{x} \) is feasible to (6.9) with \( y = \hat{y} \). Hence, using this in (6.11), we obtain \( \phi(\hat{x}, \bar{u}, \bar{V}) = f(\hat{x}) + \bar{u}'g(\hat{x}) + \bar{V}'h(\bar{x}) \) all \( \hat{x} \in X \); so \( \phi(\bar{x}, \bar{u}, \bar{V}) = \min \{ \phi(x, \bar{u}, \bar{V}) : x \in X \} \).

We have therefore shown that \( \bar{x} \in X \), \( \bar{u} \geq 0 \), and that conditions (a), (b), and (c) of Theorem 6.2.5 hold true. Consequently, \((\bar{x}, \bar{u}, \bar{V})\) is a saddle point for \( \phi \), and this completes the proof.

To illustrate, observe in Figures 6.1 and 6.2 that there does exist a supporting hyperplane for the epigraph of \( v \) at \((0, v(0))\). Hence, both the primal and dual problems have optimal solutions having the same optimal objective values for these cases. However, for the situations illustrated in Figures 6.3 and 6.5, no such supporting hyperplane exists. Hence, these instances possess a positive duality gap.
In conclusion, there are two noteworthy points regarding the perturbation function \( v \). First, if \( f \) and \( g \) are convex, \( h \) is affine, and \( X \) is a convex set, it can easily be shown that \( v \) is a convex function (see Exercise 6.4). Hence, in this case, the condition (6.10) reduces to the statement that \(- (\bar{u}', \bar{v}')\) is a subgradient of \( v \) at \( y = 0 \).

Second, suppose that corresponding to the primal and dual problems \( P \) and \( D \) there exists a saddle point solution \((\bar{x}, \bar{u}, \bar{v})\), and assume that \( v \) is continuously differentiable at \( y = 0 \). Then we have \( v(y) = v(0) + \nabla v(0)'y + \|y\|\alpha(0; y) \), where \( \alpha(0; y) \rightarrow 0 \) as \( y \rightarrow 0 \). Using (6.10) of Theorem 6.2.7, this means that \(-[\nabla v(0)' + (\bar{u}', \bar{v}')]y \leq \|y\|\alpha(0; y) \) for all \( y \in \mathbb{R}^{m+\ell} \). Letting \( y = -\lambda[\nabla v(0)' + (\bar{u}', \bar{v}')] \) for \( \lambda \geq 0 \), and letting \( \lambda \rightarrow 0^+ \), we readily conclude that \( \nabla v(0)' = -(\bar{u}', \bar{v}') \). Hence, the negative of the optimal Lagrange multiplier values give the marginal rates of change in the optimal objective value of Problem \( P \) with respect to perturbations in the right-hand sides. Assuming that the problem represents one of minimizing cost subject to various material, labor, budgetary resource limitations, and demand constraints, this yields useful economic interpretations in terms of the marginal change in the optimal cost with respect to perturbations in such resource or demand entities.

6.2.8 Example

Consider the following (primal) problem:

\[
P: \text{Minimize}\{x_2 : x_1 \geq 1, \ x_1^2 + x_2^2 \leq 1, (x_1, x_2) \in \mathbb{R}^2\}.
\]

As illustrated in Figure 6.6, the unique optimal solution to this problem is \((\bar{x}_1, \bar{x}_2) = (1, 0)\), with optimal objective function value equal to 0. However, although this is a convex programming problem, the optimum is not a KKT point, since \( F_0 \cap G' \neq \emptyset \), and a saddle point solution does not exist (refer to Theorems 4.2.15 and 6.2.6).

Now let us formulate the Lagrangian dual Problem \( D \) by treating \( 1 - x_1 \leq 0 \) as \( g(x) \leq 0 \) and letting \( X \) denote the set \( \{(x_1, x_2) : x_1^2 + x_2^2 \leq 1\} \). Hence, Problem \( D \) requires us to find \( \sup\{\theta(u) : u \geq 0\} \), where \( \theta(u) = \inf\{x_2 + u(1 - x_1) : x_1^2 + x_2^2 \leq 1\} \). For any \( u \geq 0 \), it is readily verified that the optimum is attained at \( x_1 = u/\sqrt{1+u^2} \) and \( x_2 = -1/\sqrt{1+u^2} \). Hence, \( \theta(u) = u - \sqrt{1+u^2} \). We see that as \( u \rightarrow \infty \), \( \theta(u) \rightarrow 0 \), the optimal primal objective value. Hence, \( \sup\{\theta(u) : u \geq 0\} = 0 \), but this is not attained for any \( \bar{u} \geq 0 \); that is, a maximizing solution \( \bar{u} \) does not exist.
Next, let us determine the perturbation function $v(y)$ for $y \in \mathbb{R}$. Note that $v(y) = \min\{x_2 : 1 - x_1 \leq y, \ x_1^2 + x_2^2 \leq 1\}$. Hence, we obtain $v(y) = \infty$ for $y < 0$, $v(y) = -\sqrt{y(2 - y)}$ for $0 \leq y \leq 1$, and $v(y) = -1$ for $y > 1$. This is illustrated in Figure 6.6c. Observe that there does not exist any supporting hyperplane at $(0, 0)$ for the epigraph of $v(y)$, $y \in \mathbb{R}$, since the right-hand derivative of $v$ with respect to $y$ at $y = 0$ is $-\infty$.

### 6.3 Properties of the Dual Function

In Section 6.2 we studied the relationships between the primal and dual problems. Under certain conditions, Theorems 6.2.4 and 6.2.5 showed that the optimal objective values of the primal and dual problems are equal, and hence it would be possible to solve the primal problem indirectly by solving the dual problem. To facilitate solution of the dual problem, we need to examine the properties of the dual function. In particular, we show that $\theta$ is concave, discuss its differentiability and subdifferentiability properties, and characterize its ascent and steepest ascent directions.

Throughout the rest of this chapter, we assume that the set $X$ is compact. This will simplify the proofs of several of the theorems. Note that this assumption is not unduly restrictive, since if $X$ were not bounded, we could add suitable lower and upper bounds on the variables such that the feasible region is not affected in the relative vicinity of an optimum. For convenience, we also combine the vectors $u$ and $v$ as $w$ and the functions $g$ and $h$ as $\beta$. Theorem 6.3.1 shows that $\theta$ is concave.

#### 6.3.1 Theorem

Let $X$ be a nonempty compact set in $\mathbb{R}^n$, and let $f : \mathbb{R}^n \to \mathbb{R}$ and $\beta : \mathbb{R}^n \to \mathbb{R}^{m+\ell}$ be continuous. Then $\theta$, defined by
Lagrangian Dualify and Saddle Point Optimality Conditions

\[ \theta(w) = \inf \{ f(x) + w^T \beta(x) : x \in X \}, \]

is concave over \( \mathbb{R}^{m+\ell} \).

**Proof**

Since \( f \) and \( \beta \) are continuous and \( X \) is compact, \( \theta \) is finite everywhere on \( \mathbb{R}^{m+\ell} \). Let \( w_1, w_2 \in \mathbb{R}^{m+\ell} \), and let \( \lambda \in (0,1) \). We then have

\[
\begin{align*}
\theta[\lambda w_1 + (1-\lambda)w_2] &= \inf \{ f(x) + [\lambda w_1 + (1-\lambda)w_2]^T \beta(x) : x \in X \} \\
&= \inf \{ \lambda [ f(x) + w_1^T \beta(x) ] + (1-\lambda)[ f(x) + w_2^T \beta(x) ] : x \in X \} \\
&\geq \lambda \inf \{ f(x) + w_1^T \beta(x) : x \in X \} \\
&+ (1-\lambda) \inf \{ f(x) + w_2^T \beta(x) : x \in X \} \\
&= \lambda \theta(w_1) + (1-\lambda)\theta(w_2).
\end{align*}
\]

Thus, \( \theta \) is concave, and the proof is complete.

Since \( \theta \) is concave, by Theorem 3.4.2, a local optimum of \( \theta \) is also a global optimum. This makes the maximization of \( \theta \) an attractive proposition. However, the main difficulty in solving the dual problem is that the dual function is not explicitly available, since \( \theta \) can be evaluated at a point only after a minimization subproblem is solved. In the remainder of this section we study differentiability and subdifferentiability properties of the dual function. These properties will aid us in maximizing the dual function.

**Differentiability of \( \theta \)**

We now address the question of differentiability of \( \theta \), defined by \( \theta(w) = \inf \{ f(x) + w^T \beta(x) : x \in X \} \). It will be convenient to introduce the following set of optimal solutions to the Lagrangian dual subproblem:

\[ X(w) = \{ y : y \text{ minimizes } f(x) + w^T \beta(x) \text{ over } x \in X \}. \]

The differentiability of \( \theta \) at any given point \( \bar{w} \) depends on the elements of \( X(\bar{w}) \). In particular, if the set \( X(\bar{w}) \) is a singleton, Theorem 6.3.3 shows that \( \theta \) is differentiable at \( \bar{w} \). First, however, the following lemma is needed.

**6.3.2 Lemma**

Let \( X \) be a nonempty compact set in \( \mathbb{R}^n \), and let \( f : \mathbb{R}^n \to \mathbb{R} \) and \( \beta : \mathbb{R}^n \to \mathbb{R}^{m+\ell} \) be continuous. Let \( \bar{w} \in \mathbb{R}^{m+\ell} \), and suppose that \( X(\bar{w}) \) is the singleton \( \{\bar{x}\} \). Suppose that \( w_k \to \bar{w} \), and let \( x_k \in X(w_k) \) for each \( k \). Then \( x_k \to \bar{x} \).
Proof

By contradiction, suppose that \( w_k \to \bar{w} \), \( x_k \in X(w_k) \), and \( \|x_k - \bar{x}\| > \varepsilon > 0 \) for all \( k \in \mathcal{K} \) where \( \mathcal{K} \) is some index set. Since \( X \) is compact, the sequence \( \{x_k\}_{\mathcal{K}} \) has a convergent subsequence \( \{x_{k_j}\}_{\mathcal{K}'} \), with limit \( y \) in \( X \).

Note that \( \|y - \bar{x}\| \geq \varepsilon > 0 \), and hence, \( y \) and \( \bar{x} \) are distinct. Furthermore, for each \( w_k \) with \( k \in \mathcal{K}' \), we have

\[
\sum_{k} f(x_k) + w_k^t \beta(x_k) \leq f(\bar{x}) + \sum_{k} w_k^t \beta(\bar{x}).
\]

Taking the limit as \( k \) in \( \mathcal{K}' \) approaches \( \infty \), and noting that \( x_k \to y \), \( w_k \to \bar{w} \), and that \( f \) and \( \beta \) are continuous, it follows that

\[
f(y) + \bar{w}^t \beta(y) \leq f(\bar{x}) + \bar{w}^t \beta(\bar{x}).
\]

Therefore, \( y \in X(\bar{w}) \), contradicting the assumption that \( X(\bar{w}) \) is a singleton. This completes the proof.

6.3.3 Theorem

Let \( X \) be a nonempty compact set in \( \mathbb{R}^n \), and let \( f: \mathbb{R}^n \to \mathbb{R} \), and \( \beta: \mathbb{R}^n \to \mathbb{R}^m \) be continuous. Let \( \bar{w} \in \mathbb{R}^m \) and suppose that \( X(\bar{w}) \) is the singleton \( \{\bar{x}\} \). Then \( \theta \) is differentiable at \( \bar{w} \) with gradient \( \nabla \theta(\bar{w}) = \beta(\bar{x}) \).

Proof

Since \( f \) and \( \beta \) are continuous and \( X \) is compact, for any given \( w \) there exists an \( x_w \in X(w) \). From the definition of \( \theta \), the following two inequalities hold true:

\[
\theta(w) - \theta(\bar{w}) \leq f(\bar{x}) + w^t \beta(\bar{x}) - f(\bar{x}) - \bar{w}^t \beta(\bar{x}) = (w - \bar{w})^t \beta(\bar{x}). \tag{6.12}
\]

\[
\theta(\bar{w}) - \theta(w) \leq f(x_w) + \bar{w}^t \beta(x_w) - f(x_w) - w^t \beta(x_w) = (\bar{w} - w)^t \beta(x_w). \tag{6.13}
\]

From (6.12) and (6.13) and the Schwartz inequality, it follows that

\[
0 \geq \theta(w) - \theta(\bar{w}) - (w - \bar{w})^t \beta(\bar{x}) \geq (w - \bar{w})^t [\beta(x_w) - \beta(\bar{x})]
\geq -\|w - \bar{w}\|\|\beta(x_w) - \beta(\bar{x})\|.
\]

This further implies that

\[
0 \geq \frac{\theta(w) - \theta(\bar{w}) - (w - \bar{w})^t \beta(\bar{x})}{\|w - \bar{w}\|} \geq -\|\beta(x_w) - \beta(\bar{x})\|. \tag{6.14}
\]
As \( w \to \bar{w} \), then, by Lemma 6.3.2, \( x_w \to \bar{x} \) and by the continuity of \( \beta \), \( \beta(x_w) \to \beta(\bar{x}) \). Therefore, from (6.14) we get

\[
\lim_{w \to \bar{w}} \frac{\theta(w) - \theta(\bar{w}) - (w - \bar{w})^t \beta(\bar{x})}{\|w - \bar{w}\|} = 0.
\]

Hence, \( \theta \) is differentiable at \( \bar{w} \) with gradient \( \beta(\bar{x}) \). This completes the proof.

**Subgradients of \( \theta \)**

We have shown in Theorem 6.3.1 that \( \theta \) is concave, and hence, by Theorem 3.2.5, \( \theta \) is subdifferentiable; that is, it has subgradients. As will be shown later, subgradients play an important role in the maximization of the dual function, since they lead naturally to the characterization of the directions of ascent. Theorem 6.3.4 shows that each \( \bar{x} \in X(\bar{w}) \) yields a subgradient of \( \theta \) at \( \bar{w} \).

**6.3.4 Theorem**

Let \( X \) be a nonempty compact set in \( R^n \), and let \( f: R^n \to R \) and \( \beta: R^n \to R^{m+t} \) be continuous so that for any \( \bar{w} \in R^{m+t} \), \( X(\bar{w}) \) is not empty. If \( \bar{x} \in X(\bar{w}) \), then \( \beta(\bar{x}) \) is a subgradient of \( \theta \) at \( \bar{w} \).

**Proof**

Since \( f \) and \( \beta \) are continuous and \( X \) is compact, \( X(\bar{w}) \neq \emptyset \) for any \( \bar{w} \in R^{m+t} \). Now, let \( \bar{w} \in R^{m+t} \), and let \( \bar{x} \in X(\bar{w}) \). Then

\[
\theta(w) = \inf \{ f(x) + w^t \beta(x) : x \in X \} \\
\leq f(\bar{x}) + w^t \beta(\bar{x}) \\
= f(\bar{x}) + (w - \bar{w})^t \beta(\bar{x}) + \bar{w}^t \beta(\bar{x}) \\
= \theta(\bar{w}) + (w - \bar{w})^t \beta(\bar{x}).
\]

Therefore, \( \beta(\bar{x}) \) is a subgradient of \( \theta \) at \( \bar{w} \), and the proof is complete.

**6.3.5 Example**

Consider the following primal problem:

Minimize \( -x_1 - x_2 \)

subject to

\[
x_1 + 2x_2 - 3 \leq 0, \quad x_1, x_2 = 0, 1, 2, \text{ or } 3.
\]

Letting \( g(x_1, x_2) = x_1 + 2x_2 - 3 \) and \( X = \{ (x_1, x_2) : x_1, x_2 = 0, 1, 2, \text{ or } 3 \} \), the dual function is given by
\[ \theta(u) = \inf \{-x_1 - x_2 + u(x_1 + 2x_2 - 3): x_1, x_2 = 0, 1, 2, \text{ or } 3\} \]

\[
= \begin{cases} 
-6 + 6u & \text{if } 0 \leq u \leq 1/2 \\
-3 & \text{if } 1/2 \leq u \leq 1 \\
-3u & \text{if } u \geq 1.
\end{cases}
\]

We ask the reader to plot the perturbation function for this example in Exercise 6.5 and to investigate the saddle point optimality conditions. Now let \( \bar{u} = 1/2 \).

To find a subgradient of \( \theta \) at \( \bar{u} \), consider the following subproblem:

Minimize \(-x_1 - x_2 + (1/2)(x_1 + 2x_2 - 3)\)

subject to \( x_1, x_2 = 0, 1, 2, \text{ or } 3 \).

Note that the set \( X(\bar{u}) \) of optimal solutions to the above problem is \( \{(3, 0), (3, 1), (3, 2), \text{ and } (3, 3)\} \). Thus, from Theorem 6.3.4, \( g(3, 0) = 0, g(3, 1) = 2, g(3, 2) = 4, \text{ and } g(3, 3) = 6 \) are subgradients of \( \theta \) at \( \bar{u} \). Note, however, that 3/2 is also a subgradient of \( \theta \) at \( \bar{u} \), but 3/2 cannot be represented as \( g(\bar{x}) \) for any \( \bar{x} \in X(\bar{u}) \).

From the above example, it is clear that Theorem 6.3.4 gives only a sufficient characterization of subgradients. A necessary and sufficient characterization of subgradients is given in Theorem 6.3.7. First, however, the following important result is needed. The principal conclusion of this result is stated in the corollary and holds true for any arbitrary concave function \( \theta \) (see Exercise 6.6). However, our proof of Theorem 6.3.6 is specialized to exploit the structure of the Lagrangian dual function \( \theta \).

### 6.3.6 Theorem

Let \( X \) be a nonempty compact set in \( \mathbb{R}^n \), and let \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) and \( \beta: \mathbb{R}^n \rightarrow \mathbb{R}^{m+\ell} \) be continuous. Let \( \bar{w}, \bar{d} \in \mathbb{R}^{m+\ell} \). Then the directional derivative of \( \theta \) at \( \bar{w} \) in the direction \( \bar{d} \) satisfies

\[ \theta'(\bar{w}; \bar{d}) \geq \bar{d}^t \beta(\bar{x}) \quad \text{for some } \bar{x} \in X(\bar{w}). \]

**Proof**

Consider \( \bar{w} + \lambda_k \bar{d} \), where \( \lambda_k \rightarrow 0^+ \). For each \( k \) there exists an \( x_k \in X(\bar{w} + \lambda_k \bar{d}); \) and since \( X \) is compact, there exists a convergent subsequence \( \{x_k\}_{\mathcal{K}} \) having a limit \( \bar{x} \) in \( X \). Given an \( x \in X \), note that

\[ f(x) + (\bar{w} + \lambda_k \bar{d})^t \beta(x) \geq f(x_k) + (\bar{w} + \lambda_k \bar{d})^t \beta(x_k) \]

for each \( k \in \mathcal{K} \). Taking the limit as \( k \rightarrow \infty \), it follows that

\[ f(x) + \bar{w}^t \beta(x) \geq f(\bar{x}) + \bar{w}^t \beta(\bar{x}); \]
that is, $\bar{x} \in X(\bar{w})$. Furthermore, by the definition of $\theta(\bar{w} + \lambda_k d)$ and $\theta(\bar{w})$, we get

$$
\theta(\bar{w} + \lambda_k d) - \theta(\bar{w}) = f(x_k) + (\bar{w} + \lambda_k d)^t \beta(x_k) - \theta(\bar{w}) \\
\geq \lambda_k d^t \beta(x_k).
$$

The above inequality holds true for each $k \in \mathbb{R}$. Noting that $x_k \to \bar{x}$ as $k \to \infty$, we get

$$
\lim_{k \to \infty} \frac{\theta(\bar{w} + \lambda_k d) - \theta(\bar{w})}{\lambda_k} \geq d^t \beta(\bar{x}).
$$

By Lemma 3.1.5,

$$
\theta'(\bar{w}; d) = \lim_{\lambda \to 0^+} \frac{\theta(\bar{w} + \lambda d) - \theta(\bar{w})}{\lambda}
$$

exists. In view of the above inequality, the proof is complete.

**Corollary**

Let $\partial \theta(\bar{w})$ be the collection of subgradients of $\theta$ at $\bar{w}$, and suppose that the assumptions of the theorem hold true. Then

$$
\theta'(\bar{w}; d) = \inf \{d^t \xi : \xi \in \partial \theta(\bar{w}) \}.
$$

**Proof**

Let $\bar{x}$ be as specified in the theorem. By Theorem 6.3.4, $\beta(\bar{x}) \in \partial \theta(\bar{w})$; and hence Theorem 6.3.6 implies that $\theta'(\bar{w}; d) \geq \inf \{d^t \xi : \xi \in \partial \theta(\bar{w}) \}$. Now let $\xi \in \partial \theta(\bar{w})$. Since $\theta$ is concave, $\theta(\bar{w} + \lambda d) - \theta(\bar{w}) \leq \lambda d^t \xi$. Dividing by $\lambda > 0$ and taking the limit as $\lambda \to 0^+$, it follows that $\theta'(\bar{w}; d) \leq d^t \xi$. Since this is true for each $\xi \in \partial \theta(\bar{w})$, $\theta'(\bar{w}; d) = \inf \{d^t \xi : \xi \in \partial \theta(\bar{w}) \}$, and the proof is complete.

**6.3.7 Theorem**

Let $X$ be a nonempty compact set in $\mathbb{R}^n$, and let $f: \mathbb{R}^n \to \mathbb{R}$ and $\beta: \mathbb{R}^n \to \mathbb{R}^{m+\ell}$ be continuous. Then $\xi$ is a subgradient of $\theta$ at $\bar{w} \in \mathbb{R}^{m+\ell}$ if and only if $\xi$ belongs to the convex hull of $\{\beta(y) : y \in X(\bar{w})\}$.

**Proof**

Denote the set $\{\beta(y) : y \in X(\bar{w})\}$ by $\Lambda$ and its convex hull by $\text{conv}(\Lambda)$. By Theorem 6.3.4, $\Lambda \subseteq \partial \theta(\bar{w})$; and since $\partial \theta(\bar{w})$ is convex, $\text{conv}(\Lambda) \subseteq \partial \theta(\bar{w})$. 

Using the facts that $X$ is compact and $\beta$ is continuous, it can be verified that $\Lambda$ is compact. Furthermore, the convex hull of a compact set is closed. Therefore, $\text{conv}(\Lambda)$ is a closed convex set. We shall now show that $\text{conv}(\Lambda) \supseteq \partial \theta(\mathbf{w})$.

By contradiction, suppose that there is a $\xi' \in \partial \theta(\mathbf{w})$ but not in $\text{conv}(\Lambda)$. By Theorem 2.3.4 there exist a scalar $\alpha$ and a nonzero vector $d$ such that

$$d^T \beta(y) \geq \alpha \quad \text{for each } y \in X(\mathbf{w}) \quad (6.15)$$

$$d^T \xi' < \alpha. \quad (6.16)$$

By Theorem 6.3.6 there exists a $y \in X(\mathbf{w})$ such that $\theta'(\mathbf{w}; d) \geq d^T \beta(y)$; and by (6.15) we must have $\theta'(\mathbf{w}; d) \geq \alpha$. But by the corollary to Theorem 6.3.6 and by (6.16), we get

$$\theta'(\mathbf{w}; d) = \inf \{d^T \xi : \xi \in \partial \theta(\mathbf{w}) \} \leq d^T \xi' < \alpha,$$

which is a contradiction. Therefore, $\xi \in \text{conv}(\Lambda)$, and $\partial \theta(\mathbf{w}) = \text{conv}(\Lambda)$. This completes the proof.

To illustrate, consider the problem of Example 6.2.2, for which the dual function $\theta(\mathbf{v})$, $\mathbf{v} \in \mathbb{R}$, is sketched in Figure 6.4. Note that $\theta$ is differentiable (has a unique subgradient) for all $\mathbf{v}$ except for $\mathbf{v} = -1$ and $\mathbf{v} = 2$. Consider $\mathbf{v} = 2$, for example. The set $X(2)$ is given by the set of alternative optimal solutions to the problem

$$\theta(2) = \min \{3x_2 - 6 : (x_1, x_2) \in X \}.$$ 

Hence, $X(2) = \{(0, 0), (4, 0)\}$, with $\theta(2) = -6$. By Theorem 6.3.4, the subgradients of the form $\beta(x)$ for $x \in X(2)$ are $h(0, 0) = -3$ and $h(4, 0) = 1$. Observe that in Figure 6.4 these values are the slopes of the two affine segments defining the graph of $\theta$ that are incident at the point $(\mathbf{v}, \theta(\mathbf{v})) = (2, -6)$. Therefore, as in Theorem 6.3.7, the set of subgradients of $\theta$ at $\mathbf{v} = 2$, which is given by the slopes of the set of affine supports for the hypograph of $\theta$, is precisely $[-3, 1]$, the set of convex combinations of $-3$ and $1$.

For another illustration using a bivariate function $\theta$, consider the following example.

6.3.8 Example

Consider the following primal problem:
Lagrangian Duality and Saddle Point Optimality Conditions

Minimize \(-(x_1 - 4)^2 - (x_2 - 4)^2\)
subject to \[\begin{align*}
x_1 - 3 & \leq 0 \\
-x_1 + x_2 & - 2 \leq 0 \\
x_1 + x_2 & - 4 \leq 0 \\
x_1, x_2 & \geq 0.
\end{align*}\]

In this example, we let \(g_1(x_1, x_2) = x_1 - 3\), \(g_2(x_1, x_2) = -x_1 + x_2 - 2\), and \(X = \{(x_1, x_2) : x_1 + x_2 - 4 \leq 0; x_1, x_2 \geq 0\}\). Thus, the dual function is given by

\[
\theta(u_1, u_2) = \inf \{-(x_1 - 4)^2 - (x_2 - 4)^2 + u_1(x_1 - 3) + u_2(-x_1 + x_2 - 2) : x \in X\}.
\]

We utilize Theorem 6.3.7 to determine the set of subgradients of \(\theta\) at \(\bar{u} = (1, 5)^t\).

To find the set \(X(\bar{u})\), we need to solve the following problem:

Minimize \(-(x_1 - 4)^2 - (x_2 - 4)^2 - 4x_1 + 5x_2 - 13\)
subject to \[\begin{align*}
x_1 + x_2 & - 4 \leq 0 \\
x_1, x_2 & \geq 0.
\end{align*}\]

The objective function of this subproblem is concave, and by Theorem 3.4.7 it assumes its minimum over a compact polyhedral set at one of the extreme points. The polyhedral set \(X\) has three extreme points, \((0, 0)\), \((4, 0)\), and \((0, 4)\).

Noting that \(f(0, 0) = f(4, 0) = -45\) and \(f(0, 4) = -9\), it is evident that the optimal solutions of the above subproblem are \((0, 0)\) and \((4, 0)\); that is, \(X(\bar{u}) = \{(0, 0), (4, 0)\}\). By Theorem 6.3.7, the subgradients of \(\theta\) at \(\bar{u}\) are thus given by the convex combinations of \(g(0, 0)\) and \(g(4, 0)\), that is, by convex combinations of the two vectors \((-3, -2)^t\) and \((1, -6)^t\). Figure 6.7 illustrates the set of subgradients.

Ascent and Steepest Ascent Directions

The dual problem is concerned with the maximization of \(\theta\) subject to the constraint \(u \geq 0\). Given a point \(w^t = (u^t, v^t)\), we would like to investigate the directions along which \(\theta\) increases. For the sake of clarity, first consider the following definition of an ascent direction, reiterated here for convenience.

6.3.9 Definition

A vector \(d\) is called an ascent direction of \(\theta\) at \(w\) if there exists a \(\delta > 0\) such that

\[
\theta(w + \lambda d) > \theta(w) \quad \text{for each } \lambda \in (0, \delta).
\]
Note that if \( \theta \) is concave, a vector \( d \) is an ascent direction of \( \theta \) at \( w \) if and only if \( \theta'(w;d) > 0 \). Furthermore, \( \theta \) assumes its maximum at \( w \) if and only if it has no ascent directions at \( w \), that is, if and only if \( \theta'(w;d) \leq 0 \) for each \( d \).

Using the corollary to Theorem 6.3.6, it follows that a vector \( d \) is an ascent direction of \( \theta \) at \( w \) if and only if \( \inf \{ d^T \xi : \xi \in \partial \theta(w) \} > 0 \), that is, if and only if the following inequality holds true for some \( \varepsilon > 0 \).

\[
d^T \xi \geq \varepsilon > 0 \quad \text{for each} \quad \xi \in \partial \theta(w).
\]

To illustrate, consider Example 6.3.8. The collection of subgradients of \( \theta \) at the point \((1, 5)\) is illustrated in Figure 6.7. A vector \( d \) is an ascent direction of \( \theta \) if and only if \( d^T \xi \geq \varepsilon \) for each subgradient \( \xi \), where \( \varepsilon > 0 \). In other words, \( d \) is an ascent direction if it makes an angle strictly less than 90° with each subgradient. The cone of ascent directions for this example is given in Figure 6.8. In this case, note that each subgradient is an ascent direction. However, this is not necessarily the case in general.

Since \( \theta \) is to be maximized, we are interested not only in an ascent direction but also in the direction along which \( \theta \) increases at the fastest local rate.

6.3.10 Definition

A vector \( \overline{d} \) is called a direction of steepest ascent of \( \theta \) at \( w \) if

\[
\theta'(w;\overline{d}) = \max_{\|d\| \leq 1} \theta'(w;d).
\]

Theorem 6.3.11 shows that the direction of steepest ascent of the Lagrangian dual function is given by the subgradient having the smallest Euclidean norm. As evident from the proof, this result is true for any arbitrary concave function \( \theta \).
6.3.11 Theorem

Let \( X \) be a nonempty compact set in \( \mathbb{R}^n \), and let \( f \): \( \mathbb{R}^n \rightarrow \mathbb{R} \) and \( \beta \): \( \mathbb{R}^n \rightarrow \mathbb{R}^{m+\ell} \) be continuous. The direction of steepest ascent \( \overrightarrow{d} \) of \( \theta \) at \( w \) is given below, where \( \overrightarrow{\xi} \) is the subgradient in \( \partial \theta(w) \) having the smallest Euclidean norm:

\[
\overrightarrow{d} = \begin{cases} 
0 & \text{if } \overrightarrow{\xi} = 0 \\
\overrightarrow{\xi} / \|\overrightarrow{\xi}\| & \text{if } \overrightarrow{\xi} \neq 0.
\end{cases}
\]

Proof

By Definition 6.3.10 and by the corollary to Theorem 6.3.6, the steepest ascent direction can be obtained from the following expression:

\[
\max_{\|d\| \leq 1} \theta'(w; d) = \max_{\|d\| \leq 1} \inf_{\xi \in \partial \theta(w)} d^t \xi.
\]

The reader can easily verify that

\[
\max_{\|d\| \leq 1} \theta'(w; d) = \max_{\|d\| \leq 1} \inf_{\xi \in \partial \theta(w)} d^t \xi
\]

\[
\leq \inf_{\xi \in \partial \theta(w)} \max_{\|d\| \leq 1} d^t \xi
\]

\[
= \inf_{\xi \in \partial \theta(w)} \|\xi\|
\]

\[
= \|\overrightarrow{\xi}\|.
\]
If we construct a direction \( \mathbf{d} \) such that \( \theta'(\mathbf{w}; \mathbf{d}) = \| \mathbf{\bar{\xi}} \| \), then by (6.17), \( \mathbf{d} \) is the steepest ascent direction. If \( \mathbf{\bar{\xi}} = 0 \), then for \( \mathbf{d} = 0 \), we obviously have \( \theta'(\mathbf{w}; \mathbf{d}) = \| \mathbf{\bar{\xi}} \| \). Now, suppose that \( \mathbf{\bar{\xi}} \neq 0 \), and let \( \mathbf{d} = \mathbf{\bar{\xi}} / \| \mathbf{\bar{\xi}} \| \). Note that

\[
\theta'(\mathbf{w}; \mathbf{d}) = \inf \{ \mathbf{d}' \mathbf{\bar{\xi}} : \mathbf{\bar{\xi}} \in \partial \theta(\mathbf{w}) \}
= \inf \left\{ \frac{\mathbf{\bar{\xi}}'}{\| \mathbf{\bar{\xi}} \|} : \mathbf{\bar{\xi}} \in \partial \theta(\mathbf{w}) \right\}
= \frac{1}{\| \mathbf{\bar{\xi}} \|} \inf \left\{ \| \mathbf{\bar{\xi}} \|^2 + \mathbf{\bar{\xi}}'(\mathbf{\bar{\xi}} - \mathbf{\bar{\xi}}) : \mathbf{\bar{\xi}} \in \partial \theta(\mathbf{w}) \right\}
= \| \mathbf{\bar{\xi}} \| + \frac{1}{\| \mathbf{\bar{\xi}} \|} \inf \{ \mathbf{\bar{\xi}}'(\mathbf{\bar{\xi}} - \mathbf{\bar{\xi}}) : \mathbf{\bar{\xi}} \in \partial \theta(\mathbf{w}) \}. \tag{6.18}
\]

Since \( \mathbf{\bar{\xi}} \) is the shortest vector in \( \partial \theta(\mathbf{w}) \), then by Theorem 2.4.1, \( \mathbf{\bar{\xi}}'(\mathbf{\bar{\xi}} - \mathbf{\bar{\xi}}) \geq 0 \) for each \( \mathbf{\bar{\xi}} \in \partial \theta(\mathbf{w}) \). Hence, \( \inf \{ \mathbf{\bar{\xi}}'(\mathbf{\bar{\xi}} - \mathbf{\bar{\xi}}) : \mathbf{\bar{\xi}} \in \partial \theta(\mathbf{w}) \} = 0 \) is achieved at \( \mathbf{\bar{\xi}} \).

From (6.18) it then follows that \( \theta'(\mathbf{w}; \mathbf{d}) = \| \mathbf{\bar{\xi}} \| \). Thus, we have shown that the vector \( \mathbf{d} \) specified in the theorem is the direction of steepest ascent both when \( \mathbf{\bar{\xi}} = 0 \) and when \( \mathbf{\bar{\xi}} \neq 0 \). This completes the proof.

### 6.4 Formulating and Solving the Dual Problem

Given a primal Problem P to minimize \( f(\mathbf{x}) \) subject to \( g(\mathbf{x}) \leq 0 \), \( h(\mathbf{x}) = 0 \), and \( \mathbf{x} \in X \), we have defined a Lagrangian dual Problem D to maximize \( \theta(\mathbf{u}, \mathbf{v}) \) subject to \( \mathbf{u} \geq 0 \), where \( \theta(\mathbf{u}, \mathbf{v}) \) is evaluated via the (Lagrangian) subproblem \( \theta(\mathbf{u}, \mathbf{v}) = \min \{ f(\mathbf{x}) + \mathbf{u}'g(\mathbf{x}) + \mathbf{v}'h(\mathbf{x}) : \mathbf{x} \in X \} \). In formulating this dual problem, we have dualized, that is, accommodated within the Lagrangian dual objective function, the constraints \( g(\mathbf{x}) \leq 0 \) and \( h(\mathbf{x}) = 0 \), maintaining any other constraints within the set \( X \). Different formulations of the Lagrangian dual problem might dualize different sets of constraints in constructing the Lagrangian dual function. This choice must usually be a trade-off between the ease of evaluating \( \theta(\mathbf{u}, \mathbf{v}) \) for a given \( (\mathbf{u}, \mathbf{v}) \) versus the duality gap that might exist between P and D. For example, consider the linear discrete problem

**DP:** Minimize \( \mathbf{c}'\mathbf{x} \)
subject to \( \mathbf{A}\mathbf{x} = \mathbf{b} \)
\( \mathbf{D}\mathbf{x} = \mathbf{d} \)
\( \mathbf{x} \in X \), \( \tag{6.19a} \)
where \( X \) is some compact, discrete set. Let us define the Lagrangian dual problem

\[
\text{LDP: Maximize } \{\theta(\pi) : \pi \text{ unrestricted}\}, \tag{6.19b}
\]

where \( \theta(\pi) = \min\{c^\prime x + \pi^\prime (Ax - b) : Dx = d, x \in X\} \). Because of the linearity of the objective function in the latter subproblem, we equivalently have \( \theta(\pi) = \min\{c^\prime x + \pi^\prime (Ax - b) : x \in \text{conv}[x \in X : Dx = d]\} \), where \( \text{conv}\{\cdot\} \) denotes the convex hull. It readily follows (see Exercise 6.7) that the Lagrangian dual objective value will match that of the modified Problem \( \text{DP}' \) to minimize \( c^\prime x \) subject to \( Ax = b \) and \( x \in \text{conv}\{x \in X : Dx = d\} \). Noting that \( \text{DP} \) is itself equivalent to minimizing \( c^\prime x \) subject to \( x \in \text{conv}\{x \in X : Ax = b, Dx = d\} \), we surmise how the partial convex hull operation manifested in \( \text{DP}' \) can influence the duality gap.

In this spirit, we may sometimes wish to manipulate the primal problem itself into a special form before constructing a Lagrangian dual formulation to create exploitable structures for the subproblem. For example, the discrete Problem \( \text{DP} \) stated above can be written equivalently as the problem to minimize \( \{c^\prime x : Ax = b, Dy = d, x = y, x \in X, y \in Y\} \), where \( Y \) is a copy of \( X \) in which the \( x \)-variables have been replaced by a set of matching \( y \)-variables. Now we can formulate a Lagrangian dual problem:

\[
\text{LDP: Maximize } \{\overline{\theta}(\mu) : \mu \text{ unrestricted}\}, \tag{6.20}
\]

where \( \overline{\theta}(\mu) = \min\{c^\prime x + \mu^\prime (x - y) : Ax = b, Dy = d, x \in X, y \in Y\} \). Observe that this subproblem decomposes into two separable problems over the \( x \)- and \( y \)-variables, each with a possible specially exploitable structure. Moreover, it can be shown (see Exercise 6.8) that \( \max_\mu \{\overline{\theta}(\mu)\} \geq \max_\pi \theta(\pi) \), where \( \theta \) is defined in (6.19b). Hence, the Lagrangian dual formulation \( \text{LDP} \) affords a tighter representation for the primal Problem \( \text{DP} \) in the sense that it yields a smaller duality gap than does \( \text{LDP} \). Note that, as observed previously, the value of \( \text{LDP} \) matches that of the following partial convex hull representation of the problem:

\[
\text{DP: Minimize } \{c^\prime x : x \in \text{conv}\{x \in X : Ax = b\},
\]

\[
y \in \text{conv}\{y \in Y : Dy = d\}, x = y\}.
\]

The conceptual approach leading to the formulation of \( \text{LDP} \) is called a layering strategy (because of the separable layers of constraints constructed), or a Lagrangian decomposition strategy (because of the separable decomposable structures generated). Refer to the Notes and References section for further reading on this subject matter.

Returning to the dual Problem \( \text{D} \) corresponding to the primal Problem \( \text{P} \) stated in Section 6.1, the reader will recall that we have described in the preceding section several properties of this dual function. In particular, the dual
problem requires the maximization of a concave function \( \theta(u, v) \) over the simple constraint set \( \{(u, v): u \geq 0\} \). If \( \theta \) is differentiable due to the property stated in Theorem 6.3.3, then \( \nabla \theta(u, v)' = [g(x)', h(x)'] \). Various algorithms described in subsequent chapters that are applicable to maximizing differentiable concave functions can be used to solve this dual problem. These algorithms involve the generation of a suitable ascent direction \( d \), followed by a one-dimensional line search along this direction to find a new improved solution.

To illustrate one simple scheme to find an ascent direction at a point \((u, v)\), consider the following strategy. If \( \nabla \theta(u, v) \neq 0 \), then by Theorem 4.1.2 this is an ascent direction and \( \theta \) will increase by moving from \((u, v)\) along \( \nabla \theta(u, v) \). However, if some components of \( u \) are equal to zero, and any of the corresponding components of \( g(x) \) is negative, then \( u + \lambda g(x) \geq 0 \) for \( \lambda > 0 \), thus violating the nonnegativity restriction. To handle this difficulty we can use a modified or projected direction \([g(X), h(\bar{x})]\), where \( g(\bar{x}) \) is defined as

\[
g_i(\bar{x}) = \begin{cases} 
g_i(\bar{x}) & \text{if } \bar{u}_i > 0 \\
\max\{0, g_i(\bar{x})\} & \text{if } \bar{u}_i = 0. 
\end{cases}
\]

It can then be shown (see Exercise 6.9) that \([\hat{g}(\bar{x}), h(\bar{x})]\) is a feasible ascent direction of \( \theta \) at \((u, v)\). Furthermore, \([\hat{g}(\bar{x}), h(\bar{x})]\) is zero only when the dual maximum is reached. On the other hand, suppose that \( \theta \) is nondifferentiable. In this case, the set of subgradients of \( \theta \) are characterized by Theorem 6.3.7. For \( \mathbf{d} \) to be an ascent direction of \( \theta \) at \((u, v)\), noting the corollary to Theorem 6.3.6 and the concavity of \( \theta \), we must have \( \mathbf{d}^T \xi \geq \varepsilon > 0 \) for each \( \xi \in \partial \theta(u, v) \). As a preliminary idea, the following problem can then be used for finding such a direction:

Maximize \( \varepsilon \)

subject to

\[
\begin{align*}
d^T \xi &\geq \varepsilon & \text{for } \xi &\in \partial \theta(u, v) \\
d_i &\geq 0 & \text{if } u_i &= 0 \\
-1 &\leq d_i &\leq 1 & \text{for } i = 1, \ldots, m + \ell.
\end{align*}
\]

Note that the constraints \( d_i \geq 0 \) if \( u_i = 0 \) ensure that the vector \( d \) is a feasible direction, and that the normalization constraints \(-1 \leq d_i \leq 1, \forall i\), guarantee a finite solution to the problem.

The reader may note the following difficulties associated with the above direction-finding problem:

1. The set \( \partial \theta(u, v) \) and hence the constraints of the problem are not known explicitly in advance. However, Theorem 6.3.7, which fully characterizes the subgradient set, could be of use.
The set $\partial \theta(u, v)$ usually admits an infinite number of subgradients, so that we have a linear program having an infinite number of constraints. However, if $\partial \theta(u, v)$ is a compact polyhedral set, then the constraints $d^T \xi \geq \varepsilon$ for $\xi \in \partial \theta(u, v)$ could be replaced by the constraints

$$d^T \xi_j \geq \varepsilon \quad \text{for} \ j = 1, \ldots, E,$$

where $\xi_1, \ldots, \xi_E$ are the extreme points of $\partial \theta(u, v)$. Thus, in this case, the problem reduces to a regular linear program.

To alleviate some of the above problems, we could use a row generation strategy in which only a finite number (say, $\gamma$) of representatives of the constraint set $d^T \xi \geq \varepsilon$ for $\xi \in \partial \theta(u, v)$ are used, and the resulting direction $d_{\gamma}$ is tested to ascertain if it is an ascent direction. This can be done by verifying if

$$\min \{d^T \xi : \xi \in \partial \theta(u, v)\} > 0.$$ If so, $d_{\gamma}$ can be used in the line search process. If not, the foregoing subproblem yields a subgradient $\xi_{\gamma+1}$ for which $d^T_{\gamma} \xi_{\gamma+1} \leq 0$, and thus this constraint can be added to the direction-finding problem, and the operation could then be repeated.

We ask the reader to provide the details of this scheme in Exercise 6.30. However, this type of a procedure is fraught with computational difficulties, except for small problems having simple structures. In Chapter 8 we address more sophisticated and efficient subgradient-based optimization schemes that can be used to optimize $\theta$ whenever it is nondifferentiable. These procedures employ various strategies for constructing directions based on single subgradients, possibly deflected by suitable means, or based on a bundle of subgradients collected over some local neighborhood. The directions need not always be ascent directions, but ultimate convergence to an optimum is nevertheless assured. We refer the reader to Chapter 8 and its Notes and References section for further information on this topic.

We now proceed to describe in detail one particular cutting plane or outer-linearization scheme for solving the dual Problem D. The concept of this approach is important in its own right, as it constitutes a useful ingredient for many decomposition and partitioning methods.

**Cutting Plane or Outer-Linearization Method**

The methods discussed in principle above for solving the dual problem generate at each iteration a direction of motion and adopt a step size along this direction, with a view ultimately to finding the maximum for the Lagrangian dual function. We now discuss a strategy for solving the dual problem in which, at each iteration, a function that approximates the dual function is optimized.

Recall that the dual function $\theta$ is defined by

$$\theta(u, v) = \inf \{f(x) + u^T g(x) + v^T h(x) : x \in X\}.$$
Letting \( z = \theta(u, v) \), the inequality \( z \leq f(x) + u'g(x) + v'h(x) \) must hold true for each \( x \in X \). Hence, the dual problem of maximizing \( \theta(u, v) \) over \( u \geq 0 \) is equivalent to the following problem:

\[
\begin{align*}
\text{Maximize} & \quad z \\
\text{subject to} & \quad z \leq f(x) + u'g(x) + v'h(x) \quad \text{for } x \in X \\
& \quad u \geq 0.
\end{align*}
\] (6.21)

Note that the above problem is a linear program in the variables \( z, u, \) and \( v \). Unfortunately, however, the constraints are infinite in number and are not known explicitly. Suppose that we have the points \( x_1, \ldots, x_{k-1} \) in \( X \), and consider the following approximating problem:

\[
\begin{align*}
\text{Maximize} & \quad z \\
\text{subject to} & \quad z \leq f(x_j) + u'g(x_j) + v'h(x_j) \quad \text{for } j = 1, \ldots, k-1 \\
& \quad u \geq 0.
\end{align*}
\] (6.22)

The above problem is a linear program having a finite number of constraints and can be solved by the simplex method, for example. Let \( (z_k, u_k, v_k) \) be an optimal solution to this approximating problem, sometimes referred to as the master program. If this solution satisfies (6.21), then it is an optimal solution to the Lagrangian dual problem. To check whether (6.21) is satisfied, consider the following subproblem:

\[
\begin{align*}
\text{Minimize} & \quad f(x) + u'g(x) + v'h(x) \\
\text{subject to} & \quad x \in X.
\end{align*}
\]

Let \( x_k \) be an optimal solution to the above problem, so that \( \theta(u_k, v_k) = f(x_k) + u_k'g(x_k) + v_k'h(x_k) \). If \( z_k \leq (u_k, v_k) \), then \( (u_k, v_k) \) is an optimal solution to the Lagrangian dual problem. Otherwise, for \( (u, v) = (u_k, v_k) \), the inequality (6.21) is not satisfied for \( x = x_k \). Thus, we add the constraint

\[
z \leq f(x_k) + u'g(x_k) + v'h(x_k)
\]

to the constraints in (6.22), and re-solve the master linear program. Obviously, the current optimal point \( (z_k, u_k, v_k) \) contradicts this added constraint. Thus, this point is cut away, hence the name cutting plane algorithm.

**Summary of the Cutting Plane or Outer-Linearization Method**

Assume that \( f, g, \) and \( h \) are continuous and that \( X \) is compact, so that the set \( X(u, v) \) is nonempty for each \( (u, v) \).
Initialization Step  Find a point $x_0 \in X$ such that $g(x_0) \leq 0$ and $h(x_0) = 0$. Let $k = 1$, and go to the main step.

Main Step  Solve the following master program:

Maximize $z$

subject to $z \leq f(x_j) + u^T g(x_j) + v^T h(x_j)$ for $j = 0, ..., k-1$

$u \geq 0$.

Let $(z_k, u_k, v_k)$ be an optimal solution. Solve the following subproblem:

Minimize $f(x) + u^T g(x) + v^T h(x)$

subject to $x \in X$.

Let $x_k$ be an optimal point, and let $\theta(u_k, v_k) = f(x_k) + u_k^T g(x_k) + v_k^T h(x_k)$. If $z_k = \theta(u_k, v_k)$, then stop; $(u_k, v_k)$ is an optimal dual solution. Otherwise, if $z_k > \theta(u_k, v_k)$, then add the constraint $z \leq f(x_k) + u^T g(x_k) + v^T h(x_k)$ to the master program, replace $k$ by $k + 1$, and repeat the main step.

At each iteration, a cut (constraint) is added to the master problem, and hence the size of the master problem increases monotonically. In practice, if the size of the master problem becomes excessively large, all constraints that are not binding may be thrown away. Theoretically, this might not guarantee convergence, unless, for example, the dual value has strictly increased since the last time such a deletion was executed, and the set $X$ has a finite number of elements. (See Exercise 6.28; and for a general convergence theorem, see Exercises 7.21 and 7.22.) Also, note that the optimal solution values of the master problem form a nonincreasing sequence $\{z_k\}$. Since each $z_k$ is an upper bound on the optimal value of the dual problem, we may stop after iteration $k$ if $z_k - \max_{1 \leq j \leq k} \theta(u_j, v_j) < \varepsilon$, where $\varepsilon$ is a small positive number.

Interpretation as a Tangential Approximation or Outer-Linearization Technique

The foregoing algorithm for maximizing the dual function can be interpreted as a tangential approximation technique. By the definition of $\theta$, we must have

$$\theta(u, v) \leq f(x) + u^T g(x) + v^T h(x) \quad \text{for } x \in X.$$  

Thus, for any fixed $x \in X$, the hyperplane

$$\{(u, v, z) : u \in R^m, \ v \in R^l, \ z = f(x) + u^T g(x) + v^T h(x)\}$$

bounds the function $\theta$ from above.

The master problem at iteration $k$ is equivalent to solving the following problem:
Maximize \( \hat{\theta}(u, v) \)
subject to \( u \geq 0, \)

where \( \hat{\theta}(u, v) = \min \{ f(x_j) + u^t g(x_j) + v^t h(x_j) : j = 1, \ldots, k - 1 \} \). Note that \( \hat{\theta} \) is a piecewise linear function that provides an outer approximation or outer linearization for \( \theta \) by considering only \( k - 1 \) of the bounding hyperplanes.

Let the optimal solution to the master problem be \((z_k, u_k, v_k)\). Now the subproblem is solved yielding \( \theta(u_k, v_k) \) and \( x_k \). If \( z_k > \theta(u_k, v_k) \), the new constraint \( z \leq f(x_k) + u^t g(x_k) + v^t h(x_k) \) is added to the master problem, giving a new and tighter piecewise linear approximation to \( \theta \). Since \( \theta(u_k, v_k) = f(x_k) + u_k^t g(x_k) + v_k^t h(x_k) \), the hyperplane \( \{(x, u, v) : z = f(x_k) + u^t g(x_k) + v^t h(x_k)\} \) is tangential to the graph of \( \theta \) at \((z_k, u_k, v_k)\): hence, the name tangential approximation.

### 6.4.1 Example

Minimize \( (x_1 - 2)^2 + (1/4)x_2^2 \)
subject to \( x_1 - (7/2)x_2 - 1 \leq 0 \)
\[ 2x_1 + 3x_2 = 4. \]

We let \( X = \{(x_1, x_2) : 2x_1 + 3x_2 = 4\} \), so that the Lagrangian dual function is given by

\[ \theta(u) = \min \{ (x_1 - 2)^2 + (1/4)x_2^2 + u(x_1 - (7/2)x_2 - 1) : 2x_1 + 3x_2 = 4 \}. \]  

(6.23)

The cutting plane method is initialized with a feasible solution \( x_0 = (5/4, 1/2)^t \). At Step 1 of the first iteration, we solve the following problem:

Maximize \( z \)
subject to \( z \leq 5/8 - (3/2)u \)
\[ u \geq 0. \]

The optimal solution is \((z_1, u_1) = (5/8, 0)\). At Step 2 we solve (6.23) for \( u = u_1 = 0 \), yielding an optimal solution \( x_1 = (2, 0)^t \) with \( \theta(u_1) = 0 < z_1 \). Hence, more iterations are needed. A summary of the first four iterations is given in Table 6.1.

The approximating function \( \hat{\theta} \) at the end of the fourth iteration is shown by darkened lines in Figure 6.9. The reader can easily verify that the Lagrangian dual function for this problem is given by \( \theta(u) = -(5/2)u^2 + u \) and that the hyperplanes added at Iteration 2 onward are indeed tangential to the graph of \( \theta \).
Table 6.1 Summary of Computations for Example 6.4.1

<table>
<thead>
<tr>
<th>Iteration $k$</th>
<th>Constraint Added</th>
<th>Step 1 Solution</th>
<th>Step 2 Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$z \leq 5/8 - (3/2)u$</td>
<td>$(z_k, u_k)$ = $(5/8, 0)$</td>
<td>$x'_k$ = $(2, 0)$</td>
</tr>
<tr>
<td>2</td>
<td>$z \leq 0 + u$</td>
<td>$(1/4, 1/4)$</td>
<td>$(13/8, 1/4)$</td>
</tr>
<tr>
<td>3</td>
<td>$z \leq 5/32 - (1/4)u$</td>
<td>$(1/8, 1/8)$</td>
<td>$(29/16, 1/8)$</td>
</tr>
<tr>
<td>4</td>
<td>$z \leq 5/128 + (3/8)u$</td>
<td>$(7/64, 3/16)$</td>
<td>$(55/32, 3/16)$</td>
</tr>
</tbody>
</table>

at the respective points $(z_k, u_k)$. Incidentally, the dual objective function is maximized at $\bar{u} = 1/5$ with $\theta(\bar{u}) = 1/10$. Note that the sequence $\{u_k\}$ converges to the optimal point $\bar{u} = 1/5$.

6.5 Getting the Primal Solution

Thus far we have studied several properties of the dual function and described some procedures for solving the dual problem. However, our main concern is finding an optimal solution to the primal problem.

In this section we develop some theorems that will aid us in finding a solution to the primal problem as well as solutions to perturbations of the primal problem. However, for nonconvex programs, as a result of the possible presence of a duality gap, additional work is usually needed to find an optimal primal solution.

Solutions to Perturbed Primal Problems

During the course of solving the dual problem, the following problem, which is used to evaluate the function $\theta$ at $(u, v)$, is solved frequently:

\[
\begin{align*}
\text{Minimize} & \quad f(x) + u^'g(x) + v^'h(x) \\
\text{subject to} & \quad x \in X.
\end{align*}
\]

Figure 6.9 Tangential approximation of $\theta$. 
Theorem 6.5.1 shows that an optimal solution $\bar{x}$ to the above problem is also an optimal solution to a problem that is similar to the primal problem, in which some of the constraints are perturbed. Specifically, $\bar{x}$ evaluates $v[\mathbf{g}(\bar{x}), \mathbf{h}(\bar{x})]$, where $v$ is the perturbation function defined in (6.9).

6.5.1 Theorem

Let $(u, v)$ be a given vector with $u \geq 0$. Consider the problem to minimize $f(x) + u'\mathbf{g}(x) + v'\mathbf{h}(x)$ subject to $x \in X$. Let $\bar{x}$ be an optimal solution. Then $\bar{x}$ is an optimal solution to the following problem, where $I = \{i : u_i > 0\}$:

Minimize $f(x)$
subject to $g_i(x) \leq g_i(\bar{x})$ for $i \in I$
$h_i(x) = h_i(\bar{x})$ for $i = 1, \ldots, \ell$
$x \in X$.

In particular, $\bar{x}$ solves the problem to evaluate $v[\mathbf{g}(\bar{x}), \mathbf{h}(\bar{x})]$, where $v$ is the perturbation function defined in (6.9).

Proof

Let $x \in X$ be such that $h_i(x) = h_i(\bar{x})$ for $i = 1, \ldots, \ell$ and $g_i(x) \leq g_i(\bar{x})$ for $i \in I$. Note that

$$f(x) + u'\mathbf{g}(x) + v'\mathbf{h}(x) \geq f(\bar{x}) + u'\mathbf{g}(\bar{x}) + v'\mathbf{h}(\bar{x}). \tag{6.24}$$

But since $\mathbf{h}(x) = \mathbf{h}(\bar{x})$ and $u'\mathbf{g}(x) = \sum_{i \in I} u_i g_i(x) \leq \sum_{i \in I} u_i g_i(\bar{x}) = u'\mathbf{g}(\bar{x})$, we get from (6.24) that

$$f(x) + u'\mathbf{g}(\bar{x}) \geq f(x) + u'\mathbf{g}(x) \geq f(\bar{x}) + u'\mathbf{g}(\bar{x}),$$

which shows that $f(x) \geq f(\bar{x})$. Hence, $\bar{x}$ solves the problem stated in the theorem. Moreover, since this problem is a relaxation of (6.9) for $y = \bar{y}$, where $\bar{y}' = [\mathbf{g}(\bar{x})', \mathbf{h}(\bar{x})']$, and since $\bar{x}$ is feasible to (6.9) with $y = \bar{y}$, it follows that $\bar{x}$ evaluates $v(\bar{y})$. This completes the proof.

Corollary

Under the assumptions of the theorem, suppose that $\mathbf{g}(\bar{x}) \leq 0$, $\mathbf{h}(\bar{x}) = 0$, and $u'\mathbf{g}(\bar{x}) = 0$. Then $\bar{x}$ is an optimal solution to the following problem:
Minimize $f(x)$  
subject to $g_i(x) \leq 0$ for $i \in I$  
$h_i(x) = 0$ for $i = 1, \ldots, \ell$  
$x \in X$.

In particular, $\bar{x}$ is an optimal solution to the original primal problem, and $(u, v)$ is an optimal solution to the dual problem.

**Proof**

Note that $u^T g(\bar{x}) = 0$ implies that $g_i(\bar{x}) = 0$ for $i \in I$; and from the theorem, it follows that $\bar{x}$ solves the problem stated. Also, since the feasible region of the primal problem is contained in that of the above problem, and since $\bar{x}$ is a feasible solution to the primal problem, then $\bar{x}$ is an optimal solution to the primal problem. Furthermore, $f(\bar{x}) = f(\bar{x}) + u^T g(\bar{x}) + v^T h(\bar{x}) = \theta(u, v)$, so that $(u, v)$ solves the dual problem. This completes the proof.

Of course, the conditions of the above corollary coincide precisely with the saddle point optimality conditions (a), (b), and (c) of Theorem 6.2.5, implying that $(\bar{x}, u, v)$ is a saddle point and, hence, that $\bar{x}$ and $(u, v)$ solve Problems P and D, respectively. Also, elements of the proof of Theorem 6.5.1 are evident in the proof of Theorem 6.2.7. However, our purpose in highlighting Theorem 6.5.1 and its corollary is to emphasize the role played by this result in deriving heuristic primal solutions based on solving the dual problem. As seen from Theorem 6.5.1, as the dual function $\theta$ is evaluated at a given point $(u, v)$, we obtain a point $\bar{x}$ that is an optimal solution to a problem that is closely related to the original problem, in which the constraints are perturbed from $h(x) = 0$ and $g_i(x) \leq 0$ for $i = 1, \ldots, m$, to $h(x) = h(\bar{x})$ and $g_i(x) \leq g_i(\bar{x})$ for $i = 1, \ldots, m$.

In particular, during the course of solving the dual problem, suppose that for a given $(u, v)$ with $u \geq 0$, we have $\hat{x} \in X(u, v)$. Furthermore, for some $\varepsilon > 0$, suppose that $|g_i(\hat{x})| \leq \varepsilon$ for $i \in I$, $g_i(\hat{x}) \leq \varepsilon$ for $i \not\in I$, and $|h_i(\hat{x})| \leq \varepsilon$ for $i = 1, \ldots, \ell$.

Note that if $\varepsilon$ is sufficiently small, then $\hat{x}$ is near-feasible. Now, suppose that $\bar{x}$ is an optimal solution to the primal Problem P. Then, by the definition of $\theta(u, v)$,

$$f(\hat{x}) + \sum_{i \in I} u_i g_i(\hat{x}) + \sum_{i = 1}^\ell v_i h_i(\hat{x}) \leq f(\bar{x}) + \sum_{i \in I} u_i g_i(\bar{x}) + \sum_{i = 1}^\ell v_i h_i(\bar{x}) \leq f(\bar{x})$$

since $h_i(\bar{x}) = 0$, $g_i(\bar{x}) \leq 0$, and $u_i \geq 0$. The above inequality thus implies that

$$f(\hat{x}) \leq f(\bar{x}) + \varepsilon \left[ \sum_{i \in I} u_i + \sum_{i = 1}^\ell |v_i| \right].$$
Therefore, if $\varepsilon$ is sufficiently small so that $\varepsilon[\sum_{i \in I} u_i + \sum_{i=1}^{\ell} v_i]$ is small enough, then $\hat{x}$ is a near-optimal solution. In many practical problems, such a solution is often acceptable.

Note also that in the absence of a duality gap, if $\overline{x}$ and $(\overline{u}, \overline{v})$ are, respectively, optimal primal and dual solutions, then, by Theorem 6.2.5, $(\overline{x}, \overline{u}, \overline{v})$ is a saddle point. Hence, by Property (a) of Theorem 6.2.5, $\overline{x}$ minimizes $\phi(x, \overline{u}, \overline{v})$ over $x \in X$. This means that there exists an optimal solution to the primal problem among points in the set $X(\overline{u}, \overline{v})$, where $(\overline{u}, \overline{v})$ is an optimal solution to the dual problem. Of course, not any solution $\overline{x} \in X(\overline{u}, \overline{v})$ solves the primal problem unless $\overline{x}$ is feasible to $P$ and it satisfies the complementary slackness condition $\overline{u}'g(\overline{x}) = 0$.

**Generating Primal Feasible Solutions in the Convex Case**

The foregoing discussion was concerned with general, perhaps nonconvex problems. Under suitable convexity assumptions, we can easily obtain primal feasible solutions at each iteration of the dual problem by solving a linear program. In particular, suppose that we are given a point $x_0$, which is feasible to the original problem, and let the points $x_j \in X(u_j, v_j)$ for $j = 1, \ldots, k$ be generated by an arbitrary algorithm used to maximize the dual function. Theorem 6.5.2 shows that a feasible solution to the primal problem can be obtained by solving the following linear programming problem $P'$:

$$
P': \begin{align*}
\text{Minimize} & \quad \sum_{j=0}^{k} \lambda_j f(x_j) \\
\text{subject to} & \quad \sum_{j=0}^{k} \lambda_j g(x_j) \leq 0 \\
& \quad \sum_{j=0}^{k} \lambda_j h(x_j) = 0 \\
& \quad \sum_{j=0}^{k} \lambda_j = 1 \\
& \quad \lambda_j \geq 0 \quad \text{for} \quad j = 0, \ldots, k.
\end{align*}
$$

**6.5.2 Theorem**

Let $X$ be a nonempty convex set in $R^n$, let $f: R^n \to R$ and $g: R^n \to R^m$ be convex, and let $h: R^n \to R^\ell$ be affine; that is, $h$ is of the form $h(x) = Ax - b$. Let $x_0$ be an initial feasible solution to Problem $P$, and suppose that $x_j \in X(u_j, v_j)$ for $j = 1, \ldots, k$ are generated by any algorithm for solving the dual problem.
Furthermore, let \( \lambda_j \) for \( j = 0, \ldots, k \) be an optimal solution to Problem \( P' \) defined in (6.25), and let \( \bar{x}_k = \sum_{j=0}^{k} \lambda_j x_j \). Then \( \bar{x}_k \) is a feasible solution to the primal Problem \( P \). Furthermore, letting \( z_k = \sum_{j=0}^{k} \lambda_j f(x_j) \) and \( z^* = \inf \{ f(x) : x \in X, g(x) \leq 0, h(x) = 0 \} \), if \( z_k - \theta(u, v) \leq \varepsilon \) for some \( (u, v) \) with \( u \geq 0 \), then \( f(\bar{x}_k) \leq z^* + \varepsilon \).

**Proof**

Since \( X \) is convex and \( x_j \in X \) for each \( j \), \( \bar{x}_k \in X \). Since \( g \) is convex and \( h \) is affine, and noting the constraints of Problem \( P' \), \( g(\bar{x}_k) \leq 0 \) and \( h(\bar{x}_k) = 0 \). Thus, \( \bar{x}_k \) is a feasible solution to the primal problem. Now suppose that \( z_k - \theta(u, v) \leq \varepsilon \) for some \( (u, v) \) with \( u \geq 0 \). Noting the convexity of \( f \) and Theorem 6.2.1, we get

\[
 f(\bar{x}_k) \leq \sum_{j=0}^{k} \lambda_j f(x_j) = z_k \leq \theta(u, v) + \varepsilon \leq z^* + \varepsilon
\]

and the proof is complete.

At each iteration of the dual maximization problem, we can thus obtain a primal feasible solution by solving the linear programming Problem \( P' \). Even though the primal objective values \( \{ f(\bar{x}_k) \} \) of the generated primal feasible points are not necessarily decreasing, they form a sequence that is bounded from above by the nonincreasing sequence \( \{ z_k \} \).

Note that if \( z_k \) is close enough to the dual objective value evaluated at any dual feasible point \( (u, v) \), where \( u \geq 0 \), then \( \bar{x}_k \) is a near-optimal primal feasible solution. Also note that we need not solve Problem \( P' \) in the case of the cutting plane algorithm, since it is precisely the linear programming dual of the master problem stated in Step 1 of this algorithm. Thus, the optimal variables \( \bar{\lambda}_0, \ldots, \bar{\lambda}_k \) can be retrieved easily from the solution to the master problem, and \( \bar{x}_k \) can be computed as \( \sum_{j=0}^{k} \bar{\lambda}_j x_j \). It is also worth mentioning that the termination criterion \( z_k = \theta(u_k, v_k) \) in the cutting plane algorithm can be interpreted as letting \( (u, v) = (u_k, v_k) \) and \( \varepsilon = 0 \) in the Theorem 6.5.2.

To illustrate the above procedure, consider Example 6.4.1. At the end of Iteration \( k = 1 \), we have the points \( x_0 = (5/4, 1/2)^t \) and \( x_1 = (2, 0)^t \). The associated primal point \( \bar{x}_1 \) can be obtained by solving the following linear programming problem:
Minimize \((5/8)\lambda_0\)  
subject to \(-(3/2)\lambda_0 + \lambda_1 \leq 0\)  
\[\lambda_0 + \lambda_1 = 1\]  
\[\lambda_0, \lambda_1 \geq 0.\]

The optimal solution to this problem is given by \(\lambda_0 = 2/5\) and \(\lambda_1 = 3/5\). This yields a primal feasible solution  
\[\bar{x}_1 = (2/5)(5/4, 1/2)' + (3/5)(2, 0)' = (17/10, 2/10)'.\]

As pointed out earlier, the above linear program need not be solved separately to find the values of \(\lambda_0\) and \(\lambda_1\), since its dual has already been solved during the course of the cutting plane algorithm.

### 6.6 Linear and Quadratic Programs

In this section, we discuss some special cases of Lagrangian duality. In particular, we discuss briefly duality in linear and quadratic programming. For linear programming problems, we relate the Lagrangian dual to that derived in Chapter 2 (see Theorem 2.7.3 and its corollaries). In the case of quadratic programming problems, we derive the well-known Dorn's dual program via Lagrangian duality.

#### Linear Programming

Consider the following primal linear program:

\[
\text{Minimize } c'x \\
\text{subject to } Ax = b \\
x \geq 0.
\]

Letting \(X = \{x : x \geq 0\}\), the Lagrangian dual of this problem is to maximize \(\theta(v)\), where

\[
\theta(v) = \inf\{c'x + v'(b - Ax) : x \geq 0\} = v'b + \inf\{(c' - v'A)x : x \geq 0\}.
\]

Clearly,

\[
\theta(v) = \begin{cases} 
  v'b & \text{if } (c' - v'A) \geq 0 \\
  -\infty & \text{otherwise.}
\end{cases}
\]

Hence, the dual problem can be stated as follows:

\[
\text{Maximize } v'b \\
\text{subject to } A'v \leq c.
\]
Recall that this is precisely the dual problem discussed in Section 2.7. Thus, in the case of linear programs, the dual problem does not involve the primal variables. Furthermore, the dual problem itself is a linear program, and the reader can verify that the dual of the dual problem is the original primal program. Theorem 6.6.1 summarizes the relationships between the primal and dual problems as established by Theorem 2.7.3 and its three corollaries.

6.6.1 Theorem

Consider the primal and dual linear problems stated above. One of the following mutually exclusive cases will occur:

1. The primal problem admits a feasible solution and has an unbounded objective value, in which case the dual problem is infeasible.
2. The dual problem admits a feasible solution and has an unbounded objective value, in which case the primal problem is infeasible.
3. Both problems admit feasible solutions, in which case both problems have optimal solutions $\bar{x}$ and $\bar{v}$ such that $c'\bar{x} = b'\bar{v}$ and $(c' - v'A)\bar{x} = 0$.
4. Both problems are infeasible.

Proof

See Theorem 2.7.3 and its Corollaries 1 and 3.

Quadratic Programming

Consider the following quadratic programming problem:

Minimize $(1/2)x'^THx + d'x$

subject to $Ax \leq b$,

where $H$ is symmetric and positive semidefinite, so that the objective function is convex. The Lagrangian dual problem is to maximize $\theta(u)$ over $u \geq 0$, where

$$\theta(u) = \inf\{(1/2)x'^THx + d'x + u'(Ax - b) : x \in \mathbb{R}^n\}. \quad (6.26)$$

Note that for a given $u$, the function $(1/2)x'^THx + d'x + u'(Ax - b)$ is convex, so a necessary and sufficient condition for a minimum is that the gradient must vanish; that is,

$$Hx + A'u + d = 0. \quad (6.27)$$

Thus, the dual problem can be written as follows:
Maximize \[(1/2)x'Hx + d'x + u'(Ax - b)\]
subject to \[Hx + A'u = -d\] \[(6.28)\]
\[u \geq 0.\]

Now, from (6.27), we have \[d'x + u'Ax = -x'Hx.\] Substituting this into (6.28), we derive the familiar form of Dorn's dual quadratic program:

Maximize \[-(1/2)x'Hx - b'u\]
subject to \[Hx + A'u = -d\] \[(6.29)\]
\[u \geq 0.\]

Again, by Lagrangian duality, if one problem is unbounded, then the other is infeasible. Moreover, following Theorem 6.2.6, if both problems are feasible, then they both have optimal solutions having the same objective value.

We now develop an alternative form of the Lagrangian dual problem under the assumption that \(H\) is positive definite, so that \(H^{-1}\) exists. In this case, the unique solution to (6.27) is given by \[x = -H^{-1}(d + A'u).\]

Substituting in (6.26), it follows that \[\theta(u) = (1/2)u'Du + u'c - (1/2)d'H^{-1}d,\]
where \[D = -AH^{-1}A'\] and \[c = -b - AH^{-1}d.\] The dual problem is thus given by:

Maximize \[(1/2)u'Du + u'c - (1/2)d'H^{-1}d\]
subject to \[u \geq 0.\] \[(6.30)\]

The dual problem (6.30) can be solved relatively easily using the algorithms described in Chapters 8 through 11, noting that this problem simply seeks to maximize a concave quadratic function over the nonnegative orthant. (See Exercise 6.45 for a simplified scheme.)

**Exercises**

[6.1] Consider the (singly) constrained problem to minimize \(f(x)\) subject to \(g(x) \leq 0\) and \(x \in X\). Define \(G = \{(y, z) : y = g(x), z = f(x)\text{ for some } x \in X\}\), and let \(v(y) = \min\{f(x) : g(x) \leq y, x \in X\}, y \in R\), be the associated perturbation function. Show that \(v\) is the pointwise supremum over all possible nonincreasing functions whose epigraph contains \(G\).

[6.2] Consider the problem to minimize \(f(x)\) subject to \(g_1(x) \leq 0\) and \(g_2(x) \leq 0\) as illustrated in Figure 4.13. Denote \(X = \{x : g_1(x) \leq 0\}\). Sketch the perturbation function \(v(y) = \min\{f(x) : g_2(x) \leq y, x \in X\}\) and indicate the duality gap. Pro-
Lagrangian Duality and Saddle Point Optimality Conditions

vide a possible sketch for the set \( G = \{(y, z): y = g_2(x), z = f(x)\} \) for this problem.

[6.3] Let \( \phi(x, y) \) be a continuous function defined for \( x \in X \subseteq \mathbb{R}^n \) and \( y \in Y \subseteq \mathbb{R}^m \). Show that

\[
\sup_{y \in Y} \inf_{x \in X} \phi(x, y) \leq \inf_{x \in X} \sup_{y \in Y} \phi(x, y).
\]

[6.4] Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \), \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \), and \( x \in X \), and let \( \nu: \mathbb{R}^{m+\ell} \to \mathbb{R} \) be the perturbation function defined by (6.9). Assuming that \( f \) and \( g \) are convex, \( h \) is affine, and that \( X \) is a convex set, show that \( \nu \) is a convex function.

[6.5] For the problem of Example 6.3.5, sketch the perturbation function \( \nu \) defined by (6.9), and comment on the existence of a saddle point solution.

[6.6] Let \( f: \mathbb{R}^n \to \mathbb{R} \) be a concave function, and let \( \partial f(\bar{x}) \) be the subdifferential of \( f \) at any \( \bar{x} \in \mathbb{R}^n \). Show that the directional derivative of \( f \) at \( \bar{x} \) in the direction \( d \) is given by

\[
f'(\bar{x}; d) = \inf\{\xi' d : \xi \in \partial f(\bar{x})\}.
\]

What is the corresponding result if \( f \) is a convex function?

[6.7] Consider the discrete optimization Problem DP: Minimize \( \{c'x : Ax = b, Dx = d, x \in X\} \), where \( X \) is some compact discrete set, and assume that the problem is feasible. Define \( \theta(\pi) = \min\{c'x + \pi'(Ax - b) : Dx = d, x \in X\} \) for any \( \pi \in \mathbb{R}^m \), where \( A \) is \( m \times n \). Show that \( \max\{\theta(\pi) : \pi \in \mathbb{R}^m\} = \min\{c'x : Ax = b, x \in \text{conv}\{x \in X : Dx = d\}\} \), where \( \text{conv}\{\cdot\} \) denotes the convex hull operation. Use this result to interpret the duality gap that might exist between DP and the Lagrangian dual problem stated.

[6.8] Consider Problem DP given in Exercise 6.7, and rewrite this problem as minimize \( \{c'x : Ax = b, Dy = d, x = y, x \in X, y \in Y\} \), where \( Y \) is a copy of \( X \) in which the \( x \)-variables have been replaced by a set of matching \( y \)-variables. Formulate a Lagrangian dual function \( \tilde{\theta}(\mu) = \min\{c'x + \mu'(x - y) : Ax = b, Dy = d, x \in X, y \in Y\} \). Show that \( \max\{\tilde{\theta}(\mu) : \mu \in \mathbb{R}^m\} = \max\{\theta(\pi) : \pi \in \mathbb{R}^m\} \), where \( \theta \) is defined in Exercise 6.7. Discuss this result in relation to the respective partial convex hulls corresponding to \( \theta \) and \( \tilde{\theta} \) as presented in Section 6.4 and Exercise 6.7.

[6.9] Consider the pair of primal and dual Problems P and D stated in Section 6.1, and assume that the Lagrangian dual function \( \theta \) is differentiable. Given \( (\tilde{u}, \tilde{v}) \in \mathbb{R}^{m+\ell}, \tilde{u} \geq 0 \), let \( \nabla \theta(\tilde{u}, \tilde{v})' = [g(\bar{x})', h(\bar{x})'] \), and define \( \tilde{g}_i(\bar{x}) = g_i(\bar{x}) \) if \( \tilde{u}_i > 0 \) and \( \tilde{g}_i(\bar{x}) = \max\{0, g_i(\bar{x})\} \) if \( \tilde{u}_i = 0 \), for \( i = 1, \ldots, m \). If \( (d_u, d_v) = [\tilde{g}(\bar{x}), \]
h(\bar{x}) \neq (0, 0), then show that \((d_u, d_v)\) is a feasible ascent direction of \(\theta\) at \((\bar{u}, \bar{v})\). Hence, discuss how \(\theta\) can be maximized in the direction \((d_u, d_v)\) via the one-dimensional problem to maximize \(\lambda \{ \theta(\bar{u} + \lambda d_u, \bar{v} + \lambda d_v) : \bar{u} + \lambda d_u \geq 0, \lambda \geq 0 \}\). On the other hand, if \((d_u, d_v) = (0, 0)\), then show that \((\bar{u}, \bar{v})\) solves D.

Consider the problem to minimize \(x_1^2 + x_2^2\) subject to \(g_1(x) = -x_1 - x_2 + 4 \leq 0\) and \(g_2(x) = x_1 + 2x_2 - 8 \leq 0\). Illustrate the gradient method presented above by starting at the dual solution \((u_1, u_2) = (0, 0)\) and verifying that after one iteration of this method, an optimal solution is obtained in this case.

**[6.10]** Consider the problem to minimize \(x_1^2 + x_2^2\) subject to \(x_1 + x_2 - 4 \geq 0\) and \(x_1, x_2 \geq 0\).

a. Verify that the optimal solution is \(\bar{x} = (2, 2)'\) with \(f(\bar{x}) = 8\).

b. Letting \(X = \{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0\}\), write the Lagrangian dual problem. Show that the dual function is \(\theta(u) = -u^2 / 2 - 4u\). Verify that there is no duality gap for this problem.

c. Solve the dual problem by the cutting plane algorithm of Section 6.4. Start with \(x = (3, 3)'\).

d. Show that \(\theta\) is differentiable everywhere, and solve the problem using the gradient method of Exercise 6.9.

**[6.11]** Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad (x_1 - 2)^2 + (x_2 - 6)^2 \\
\text{subject to} & \quad x_1^2 - x_2 \leq 0 \\
& \quad -x_1 \leq 1 \\
& \quad 2x_1 + 3x_2 \leq 18 \\
& \quad x_1, x_2 \geq 0.
\end{align*}
\]

a. Find the optimal solution geometrically, and verify it by using the KKT conditions.

b. Formulate the dual problem in which \(X = \{(x_1, x_2) : 2x_1 + 3x_2 \leq 18, x_1, x_2 \geq 0\}\).

c. Perform three iterations of the cutting plane algorithm described in Section 6.4, starting with \((u_1, u_2) = (0, 0)\). Describe the perturbed optimization problems corresponding to the generated primal infeasible points. Also identify the primal feasible solutions generated by the algorithm.

**[6.12]** In reference to Exercise 6.11, perform three iterations of the gradient method of Exercise 6.9 and compare the results with those obtained by the cutting plane algorithm.

**[6.13]** Consider the following problem:
Maximize $3x_1 + 6x_2 + 2x_3 + 4x_4$
subject to $x_1 + x_2 + x_3 + x_4 \leq 12$
$-x_1 + x_2 + 2x_4 \leq 4$
$x_1 + x_2 \leq 12$
$x_2 \leq 4$
$x_3 + x_4 \leq 6$
$x_1, x_2, x_3, x_4 \geq 0$.

a. Formulate the dual problem in which $X = \{(x_1, x_2, x_3): x_1 + x_2 \leq 12, x_2 \leq 4, x_3 + x_4 \leq 6; x_1, x_2, x_3, x_4 \geq 0\}$.

b. Starting from the point $(0, 0)$, solve the Lagrangian dual problem by optimizing along the direction of steepest ascent as discussed in Exercise 6.9.

c. At optimality of the dual, find the optimal primal solution.

[6.14] Consider the primal Problem P discussed in Section 6.1. Introducing the slack vector $s$, the problem can be formulated as follows:

Minimize $f(x)$
subject to $g(x) + s = 0$
$h(x) = 0$
$(x, s) \in X'$,

where $X' = \{(x, s): x \in X, s \geq 0\}$. Formulate the dual of the above problem and show that it is equivalent to the dual problem discussed in Section 6.1.

[6.15] Consider the following problem:

Maximize $3x_1 + 2x_2 + x_3$
subject to $2x_1 + x_2 - x_3 \leq 2$
$x_1 + 2x_2 \leq 4$
$x_3 \leq 3$
$x_1, x_2, x_3 \geq 0$.

a. Find explicitly the dual function, where $X = \{(x_1, x_2, x_3): 2x_1 + x_2 - x_3 \leq 2; x_1, x_2, x_3 \geq 0\}$.

b. Repeat Part a for $X = \{(x_1, x_2, x_3): x_1 + 2x_2 \leq 4; x_1, x_2, x_3 \geq 0\}$.

c. In Parts a and b, note that the difficulty in evaluating the dual function at a given point depends on which constraints are handled via the set $X$. Propose some general guidelines that could be used in selecting the set $X$ to make the solution easier.

[6.16] Consider the problem to minimize $e^{-2x}$ subject to $-x \leq 0$.

a. Solve the above primal problem.
b. Letting $X = \mathbb{R}$ find the explicit form of the Lagrangian dual function, and solve the dual problem.

[6.17] Consider the problem to minimize $x_1$ subject to $x_1^2 + x_2^2 = 4$. Derive the dual function explicitly, and verify its concavity. Find the optimal solutions to both the primal and dual problems, and compare their objective values.

[6.18] Under the assumptions of Theorem 6.2.5, suppose that $\bar{x}$ is an optimal solution to the primal problem and that $f$ and $g$ are differentiable at $\bar{x}$. Show that there exists a vector $(\bar{u}, \bar{v})$ such that

$$
\left[ \nabla f(\bar{x}) + \sum_{i=1}^{m} \bar{u}_i \nabla g_i(\bar{x}) + \sum_{i=1}^{l} \bar{v}_i \nabla h_i(\bar{x}) \right]' (x - \bar{x}) \geq 0 \quad \text{for each } x \in X
$$

$$
u_i g_i(\bar{x}) = 0 \quad \text{for } i = 1, \ldots, m
$$

$$\bar{u} \geq 0.
$$

Show that these conditions reduce to the KKT conditions if $X$ is open.

[6.19] Consider the problem to minimize $f(x)$ subject to $g(x) \leq 0$, $x \in X$. Theorem 6.2.4 shows that the primal and dual objective values are equal at optimality under the assumptions that $f$, $g$, and $X$ are convex and that the constraint qualification $g(\bar{x}) < 0$ for some $\bar{x} \in X$ holds true. Suppose that the convexity assumptions on $f$ and $g$ are replaced by continuity of $f$ and $g$ and that $X$ is assumed to be convex and compact. Does the result of the theorem hold true? Prove or give a counterexample.

[6.20] In the proof of Lemma 6.2.3, show that the set $\Lambda$ is convex.

[6.21] Prove the following saddle point optimality condition. Let $X$ be a nonempty convex set in $\mathbb{R}^n$, and let $f: \mathbb{R}^n \to \mathbb{R}$, $g: \mathbb{R}^n \to \mathbb{R}^m$ be convex and $h: \mathbb{R}^n \to \mathbb{R}^l$ affine. If $\bar{x}$ is an optimal solution to the problem to minimize $f(x)$ subject to $g(x) \leq 0$, $h(x) = 0$, $x \in X$, then there exist $(\bar{u}_0, \bar{u}, \bar{v}) \neq 0$, $(\bar{u}_0, \bar{u}) \geq 0$ such that

$$
\phi(\bar{u}_0, u, v, \bar{x}) \leq \phi(\bar{u}_0, \bar{u}, \bar{v}, \bar{x}) \leq \phi(\bar{u}_0, \bar{u}, \bar{v}, x)
$$

for all $u \geq 0$, $v \in \mathbb{R}^l$ and $x \in X$, where $\phi(u_0, u, v, x) = u_0 f(x) + u' g(x) + v' h(x)$.

[6.22] Let $P$ and $D$ be the primal and dual nonlinear programs stated in Section 6.1, and denote $w = (u, v)$. Suppose that $\tilde{w}$ solves $D$. If there exists a saddle point solution to $P$ and if $\bar{x}$ solves uniquely for $\theta(\tilde{w})$, then show that $(\bar{x}, \tilde{w})$ is such a saddle point solution. Correspondingly, if $\theta$ is differentiable at $\tilde{w}$, and if $\bar{x}$ (uniquely) solves for $\theta$ at $\tilde{w}$, then show that $(\bar{x}, \tilde{w})$ is a saddle point solution. (In particular, this shows that if Problem $P$ has no saddle point solution, then $\theta$ cannot be differentiable at optimality.)
Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad -2x_1 + 2x_2 + x_3 - 3x_4 \\
\text{subject to} & \quad x_1 + x_2 + x_3 + x_4 \leq 8 \\
& \quad x_1 - 2x_3 + 4x_4 \leq 2 \\
& \quad x_1 + x_2 \leq 8 \\
& \quad x_3 + 2x_4 \leq 6 \\
& \quad x_1, x_2, x_3, x_4 \geq 0.
\end{align*}
\]

Let \( X = \{(x_1, x_2, x_3, x_4) : x_1 + x_2 \leq 8, x_3 + 2x_4 \leq 6; x_1, x_2, x_3, x_4 \geq 0\} \).

a. Find the function \( \theta \) explicitly.

b. Verify that \( \theta \) is differentiable at \((4,0)\), and find \( \nabla \theta(4,0) \).

c. Verify that \( \nabla \theta(4,0) \) is an infeasible direction, and find an improving feasible direction.

d. Starting from \((4,0)\), maximize \( \theta \) in the direction obtained in Part c.

Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad 2x_1 + x_2 \\
\text{subject to} & \quad x_1 + 2x_2 \leq 8 \\
& \quad 2x_1 + 3x_2 \leq 6 \\
& \quad x_1, x_2 \geq 0 \\
& \quad x_1, x_2 \text{ integers.}
\end{align*}
\]

Let \( X = \{(x_1, x_2) : 2x_1 + 3x_2 \leq 6, x_1, x_2 \geq 0 \text{ and integer}\} \). At \( u = 2 \), is \( \theta \) differentiable? If not, characterize its ascent directions.

Construct a numerical problem in which a subgradient of the dual function is not an ascent direction. Is it possible that the collection of subgradients and the cone of ascent directions are disjoint at a nonoptimal solution? (Hint: Consider the shortest subgradient.)

Suppose that \( \theta : \mathbb{R}^m \to \mathbb{R} \) is concave.

a. Show that \( \theta \) achieves its maximum at \( \bar{u} \) if and only if

\[
\max \{\theta'(\bar{u};d) : \|d\| \leq 1\} = 0.
\]

b. Show that \( \theta \) achieves its maximum over the region \( U = \{u : u \geq 0\} \) at \( \bar{u} \) if and only if

\[
\max \{\theta'(\bar{u};d) : d \in D, \|d\| \leq 1\} = 0,
\]

where \( D \) is the cone of feasible directions of \( U \) at \( \bar{u} \). (Note that the above results can be used as stopping criteria for maximizing the Lagrangian dual function.)
[6.27] Consider the problem to minimize $x$ subject to $g(x) \leq 0$ and $x \in X = \{x : x \geq 0\}$. Derive the explicit form of the Lagrangian dual function, and determine the collection of subgradients at $u = 0$ for each of the following cases:

a. $g(x) = \begin{cases} -2/x & \text{for } x \neq 0 \\ 0 & \text{for } x = 0. \end{cases}$

b. $g(x) = \begin{cases} -2/x & \text{for } x \neq 0 \\ -1 & \text{for } x = 0. \end{cases}$

c. $g(x) = \begin{cases} 2/x & \text{for } x \neq 0 \\ 1 & \text{for } x = 0. \end{cases}$

[6.28] Consider the cutting plane method described in Section 6.4, and suppose that each time the master program objective value strictly increases, we delete all the constraints of the type $z \leq f(x_i) + u^T g(x_j) + v^T h(x_j)$ that are nonbinding at optimality. If $X$ has a finite number of elements, show that this modified algorithm will converge finitely. Give some alternative conditions under which such a constraint deletion will assure convergence of the algorithm.

[6.29] Consider the following problem, in which $X$ is a compact polyhedral set and $f$ is a concave function:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad Ax = b \\
& \quad x \in X.
\end{align*}
\]

a. Formulate the Lagrangian dual problem.

b. Show that the dual function is concave and piecewise linear.

c. Characterize the subgradients, the ascent directions, and the steepest ascent direction for the dual function.

d. Generalize the result in Part b to the case where $X$ is not compact.

[6.30] Consider the pair of primal and dual Problems P and D stated in Section 6.1, and suppose that the Lagrangian dual function $\theta$ is not necessarily differentiable. Given $\bar{w} = (\bar{u}, \bar{v}) \in R^{m+\ell}$, $\bar{u} \geq 0$, let $\xi_1, \ldots, \xi_p$, $p \geq 1$, be some known collection of subgradients of $\theta$ at $\bar{w}$. Consider the problem to maximize \( \{\varepsilon : d_i^T \xi_j \geq \varepsilon \text{ for } j = 1, \ldots, p, -1 \leq d_i \leq 1 \text{ for } i = 1, \ldots, m+\ell, \text{ with } d_i \geq 0 \text{ if } \bar{u}_i = 0\} \). Let $(\varepsilon, \bar{d})$ solve this problem. If $\varepsilon = 0$, show that $\bar{w}$ solves D. Otherwise, solve the problem to maximize $\{\bar{d}_i^T \xi : \xi \in \partial \theta(\bar{w})\}$, and let $\xi_{p+1}$ be an optimum. If $\bar{d}_i^T \xi_{p+1} > 0$, then show that $\bar{d}$ is an ascent direction along which $\theta$ can be maximized by solving $\max\{\theta(\bar{w} + \lambda \bar{d}) : \bar{u}_i + \lambda \bar{d}_i \geq 0 \text{ for } i = 1, \ldots, m, \lambda \geq 0\}$, and the process can then be repeated. Otherwise, if $\bar{d}_i^T \xi_{p+1} \leq 0$, then increment $p$ by 1 and re-solve the direction-finding problem given above. Discuss the possible
computational difficulties associated with this scheme. How would you implement the various steps if all functions were affine and \( X \) was a nonempty polytope? Illustrate this using the example to minimize \( x_1 - 4x_2 \) subject to 
\[-x_1 - x_2 + 2 \leq 0, \ x_2 - 1 \leq 0 \text{ and } x \in X = \{ x : 0 \leq x_1 \leq 3, 0 \leq x_2 \leq 3 \}, \]
starting at the point \((u_1, u_2) = (0, 4)\).

[6.31] Consider the linear program to minimize \( c^T x \) subject to \( Ax = b, x \geq 0 \). Write the dual problem. Show that the dual of the dual problem is equivalent to the primal problem.

[6.32] Consider the following problem:

Minimize \(-2x_1 - 2x_2 - x_3\)
subject to 
\[
\begin{align*}
2x_1 + x_2 + x_3 &\leq 8 \\
3x_1 - 2x_2 + 3x_3 &\leq 3 \\
x_1 + x_2 &\leq 5 \\
x_1, x_2, x_3 &\geq 0.
\end{align*}
\]

Solve the primal problem by the simplex method. At each iteration identify the dual variables from the simplex tableau. Show that the dual variables satisfy the complementary slackness conditions but violate the dual constraints. Verify that dual feasibility is attained at termination.

[6.33] Consider the primal and dual linear programming problems discussed in Section 6.6. Show directly using Farkas's lemma that if the primal is inconsistent and the dual admits a feasible solution, the dual has an unbounded objective value.

[6.34] In Section 6.3 we showed that the shortest subgradient \( \xi \) of \( \theta \) at \( \overline{u} \) is the steepest ascent direction. The following modification of \( \xi \) is proposed to maintain feasibility:

\[
\overline{\xi}_i = \begin{cases} 
\max\{0, \xi_i\} & \text{if } \overline{u}_i = 0 \\
\xi_i & \text{if } \overline{u}_i \geq 0.
\end{cases}
\]

Is \( \overline{\xi} \) an ascent direction? Is it the direction of steepest ascent with the added nonnegativity restriction? Prove or give a counterexample.

[6.35] Suppose that the shortest subgradient \( \overline{\xi} \) of \( \theta \) at \( (\overline{u}, \overline{v}) \) is not equal to zero. Show that there exists an \( \varepsilon > 0 \) such that \( \|\xi - \overline{\xi}\| < \varepsilon \) implies that \( \xi \) is an ascent direction of \( \theta \) at \( (\overline{u}, \overline{v}) \). (From this exercise, if an iterative procedure is used to find \( \xi \), it would find an ascent direction after a sufficient number of iterations.)
Consider a singly constrained problem to minimize \( f(x) \) subject to \( g(x) \leq 0 \) and \( x \in X \), where \( X \) is a compact set. The Lagrangian dual problem is to maximize \( \theta(u) \) subject to \( u \geq 0 \), where \( \theta(u) = \inf \{ f(x) + ug(x) : x \in X \} \).

(a) Let \( \hat{u} \geq 0 \), and let \( \hat{x} \in X(\hat{u}) \). Show that if \( g(\hat{x}) > 0 \), then \( \bar{u} > \hat{u} \), and if \( g(\hat{x}) < 0 \), then \( \bar{u} < \hat{u} \), where \( \bar{u} \) is an optimal solution to the Lagrangian dual.

(b) Use the result of Part a to find an interval \([a, b]\) that contains all the optimal solutions to the dual problem or else to conclude that the dual problem is unbounded.

(c) Now consider the problem to maximize \( \theta(u) \) subject to \( a \leq u \leq b \).

Suppose that the following scheme is used to solve the problem: Let \( \bar{u} = (a + b)/2 \), and let \( \bar{x} \in X(\bar{u}) \). If \( g(\bar{x}) > 0 \), replace \( a \) by \( \bar{u} \) and repeat the process. If \( g(\bar{x}) < 0 \), replace \( b \) by \( \bar{u} \) and repeat the process. If \( g(\bar{x}) = 0 \), stop; \( \bar{u} \) is an optimal dual solution. Show that the procedure converges to an optimal solution, and illustrate by solving the dual of the following problem:

\[
\begin{align*}
\text{Minimize} & \quad 2x_1^2 + x_2^2 \\
\text{subject to} & \quad -x_1 - 2x_2 + 2 \leq 0.
\end{align*}
\]

(d) An alternative approach to solving the problem to maximize \( \theta(u) \) subject to \( a \leq u \leq b \) is to specialize the tangential approximation method discussed in Section 6.4. Show that at each iteration only two supporting hyperplanes need be considered, and that the method could be stated as follows: Let \( x_a \in X(a) \) and \( x_b \in X(b) \). Let \( \bar{u} = [f(x_a) - f(x_b)]/[g(x_b) - g(x_a)] \). If \( \bar{u} = a \) or \( \bar{u} = b \), stop; \( \bar{u} \) is an optimal solution to the dual problem. Otherwise, let \( \bar{x} \in X(\bar{u}) \). If \( g(\bar{x}) > 0 \), replace \( a \) by \( \bar{u} \) and repeat the process. If \( g(\bar{x}) < 0 \), replace \( b \) by \( \bar{u} \) and repeat the process. If \( g(\bar{x}) = 0 \), stop; \( \bar{u} \) is an optimal dual solution. Show that the procedure converges to an optimal solution, and illustrate by solving the problem in Part c.

Consider the primal and Lagrangian dual problems discussed in Section 6.1. Let \((\bar{u}, \bar{v})\) be an optimal solution to the dual problem. Given \((u, v)\), suppose that \( \bar{x} \in X(u, v) \), as defined in Section 6.3. Show that there exists a \( \delta > 0 \) such that \( \| (\bar{u}, \bar{v}) - (u, v) - \lambda [g(\bar{x}), h(\bar{x})] \| \) is a nonincreasing function of \( \lambda \) over the interval \([0, \delta]\). Interpret the result geometrically, and illustrate by the following problem, in which \((u_1, u_2) = (3, 1)\) are the dual variables corresponding to the first two constraints:
Minimize \(-2x_1 - 2x_2 - 5x_3\)
subject to \(x_1 + x_2 + x_3 \leq 10\)
\(x_1 + 2x_3 \geq 6\)
\(x_1, x_2, x_3 \leq 3\)
\(x_1, x_2, x_3 \geq 0\).

[6.38] From Exercise 6.37 it is clear that moving a small step in the direction of any subgradient leads us closer to an optimal dual solution. Consider the following algorithm for maximizing the dual of the problem to minimize \(f(x)\) subject to \(h(x) = 0, x \in X\).

**Main Step**

Given \(v_k\), let \(x_k \in X(v_k)\). Let \(v_{k+1} = v_k + \lambda h(x_k)\), where \(\lambda > 0\) is a small scalar. Replace \(k\) by \(k + 1\) and repeat the main step.

a. Discuss some possible ways of choosing a suitable step size \(\lambda\). Do you see any advantages in reducing the step size during later iterations? If so, propose a scheme for doing that.

b. Does the dual function necessarily increase from one iteration to another? Discuss.

c. Devise a suitable termination criterion.

d. Apply the above algorithm, starting from \(v = (1, 2)'\), to solve the following problem:

Minimize \(x_1^2 + x_2^2 + 2x_3\)
subject to \(x_1 + x_2 + x_3 = 6\)
\(-x_1 + x_2 + x_3 = 4\).

(This procedure, with a suitable step size selection rule, is referred to as a subgradient optimization technique. See Chapter 8 for further details.)

[6.39] Consider the problem to minimize \(f(x)\) subject to \(g(x) \leq 0, x \in X\).

a. In Exercise 6.38, a subgradient optimization technique was discussed for the equality case. Modify the procedure for the above inequality-constrained problem. \([Hint: \text{Given } u, \text{let } x \in X(u). \text{Replace } g_i(x) \text{ by } \max\{0, g_i(x)\} \text{ for each } i \text{ with } u_i = 0.]\)

b. Illustrate the procedure given in Part a by solving the problem in Exercise 6.13 starting from \(u = (0, 0)'\).

c. Extend the subgradient optimization technique to handle both equality and inequality constraints.

[6.40] Consider the problems to find
\[
\min \max_{x \in X} \phi(x, y) \quad \text{and} \quad \max_{y \in Y} \min_{x \in X} \phi(x, y),
\]

where \(X\) and \(Y\) are nonempty compact convex sets in \(\mathbb{R}^n\) and \(\mathbb{R}^m\), respectively, and \(\phi\) is convex in \(x\) for any given \(y\) and concave in \(y\) for any given \(x\).

a. Show that \(\min \max_{x \in X} \phi(x, y) \geq \max \min_{y \in Y} \phi(x, y)\) without any convexity assumptions.

b. Show that \(\max_{y \in Y} \phi(\cdot, y)\) is a convex function in \(x\) and that \(\min_{x \in X} \phi(x, \cdot)\) is a concave function in \(y\).

c. Show that \(\min \max_{x \in X} \phi(x, y) = \max \min_{y \in Y} \phi(x, y)\).

\(\text{(Hint: Use Part b and the necessary optimality conditions of Section 3.4.)}\)

[6.41] Consider the following problem, in which \(X\) is a compact polyhedral set:

\[
\begin{align*}
\text{Minimize} \quad & c'x \\
\text{subject to} \quad & Ax = b \\
& x \in X.
\end{align*}
\]

For a given dual vector \(v\), suppose that \(x_1, \ldots, x_k\) are the extreme points in \(X\) that belong to \(X(v)\) as defined in Section 6.3. Show that the extreme points of \(\partial \theta(v)\) are contained in the set \(\Lambda = \{Ax_j - b : j = 1, \ldots, k\}\). Give an example where the extreme points of \(\partial \theta(v)\) form a proper subset of \(\Lambda\).

[6.42] A company wants to plan its production rate of a certain item over the planning period \([0, T]\) such that the sum of its production and inventory costs is minimized. In addition, the known demand must be met, the production rate must fall in the acceptable interval \([\ell, u]\), the inventory must not exceed \(d\), and it must be at least equal to \(b\) at the end of the planning period. The problem can be formulated as follows:

\[
\begin{align*}
\text{Minimize} \quad & \int_0^T [c_1x(t) + c_2y^2(t)] \, dt \\
\text{subject to} \quad & x(t) = x_0 + \int_0^t [y(\tau) - z(\tau)] \, d\tau \quad \text{for } t \in [0, T] \\
& x(T) \geq b \\
& 0 \leq x(t) \leq d \quad \text{for } t \in (0, T) \\
& \ell \leq y(t) \leq u \quad \text{for } t \in (0, T),
\end{align*}
\]

where

\[
\begin{align*}
x(t) &= \text{inventory at time } t \\
y(t) &= \text{production rate at time } t \\
z(t) &= \text{known demand rate at time } t
\end{align*}
\]
Lagrangian Duality and Saddle Point Optimality Conditions

\[x_0 = \text{known initial inventory}\]
\[c_1, c_2 = \text{known coefficients}\]

a. Make the above control problem discrete as was done in Section 1.2, and formulate a suitable Lagrangian dual problem.

b. Make use of the results of this chapter to develop a scheme for solving the primal and dual problems.

c. Apply your algorithm to the following data: \(T = 6, x_0 = 0, b = 4, c_1 = 1, c_2 = 2, \ell = 2, u = 5, d = 6,\) and \(z(t) = 4\) over \([0, 4]\) and \(z(t) = 3\) over \((4, 6]\).

[6.43] Consider the following warehouse location problem. We are given destinations \(1, \ldots, k,\) where the known demand for a certain product at destination \(j\) is \(d_j.\) We are also given \(m\) possible sites for building warehouses. If we decide to build a warehouse at site \(i,\) its capacity has to be \(s_i,\) and it incurs a fixed cost \(f_i.\) The unit shipping cost from warehouse \(i\) to destination \(j\) is \(c_{ij}.\) The problem is to determine how many warehouses to build, where to locate them, and what shipping patterns to use so that the demand is satisfied and the total cost is minimized. The problem can be stated mathematically as follows:

\[
\begin{align*}
\text{Minimize} & \quad \sum_{i=1}^{m} \sum_{j=1}^{k} c_{ij}x_{ij} + \sum_{i=1}^{m} f_iy_i \\
\text{subject to} & \quad \sum_{j=1}^{k} x_{ij} \leq s_iy_i \quad \text{for } i = 1, \ldots, m \\
& \quad \sum_{i=1}^{m} x_{ij} \geq d_j \quad \text{for } j = 1, \ldots, k \\
& \quad 0 \leq x_{ij} \leq y_i \min\{s_i, d_j\} \quad \text{for } i = 1, \ldots, m; j = 1, \ldots, k \\
& \quad y_i = 0 \text{ or } 1 \quad \text{for } i = 1, \ldots, m.
\end{align*}
\]

a. Formulate a suitable Lagrangian dual problem. Explain the utility of the upper bound imposed on \(x_{ij}.\)

b. Make use of the results of this chapter to devise a special scheme for maximizing the dual of the warehouse location problem.

c. Illustrate by a small numerical example.

[6.44] Consider the (primal) quadratic program PQP: Minimize \(c^T x + (1/2) x^T Dx : Ax \geq b\), where \(D\) is an \(n \times n\) symmetric matrix and \(A\) is \(m \times n.\) Let \(W\) be an arbitrary set such that \(\{w : Aw \geq b\} \subseteq W,\) and consider Problem EDQP: Minimize \(c^T x + (1/2) w^T Dw : Ax \geq b, Dw = Dx, w \in W\).

a. Show that PQP and EDQP are equivalent in the sense that if \(x\) is feasible to PQP, \((x, w)\) with \(w = x\) is feasible to EDQP with the
same objective value; and conversely, if \((x, w)\) is feasible to EPQP, \(x\) is feasible to PQP with the same objective value.

b. Construct a Lagrangian dual \(LD\): Maximize \(\{\theta(y)\}\), where \(\theta(y) = \min\{(c + Dy)'x + (1/2)w'Dw - y'Dw : Ax \geq b, w \in W\}\). Show that equivalently, we have

\[
LD: \sup\{b'u - (1/2)y'Dy + \phi(y) : A'u - Dy = c, u \geq 0\},
\]

where \(\phi(y) = \inf\{(1/2)(y - w)'D(y - w) : w \in W\}\).

c. Show that if \(D\) is positive semidefinite and \(W = R^n\), \(\phi(y) = 0\) for all \(y\), and \(LD\) reduces to Dorn’s dual program given in (6.29). On the other hand, if \(D\) is not positive semidefinite and \(W = R^n\), \(\phi(y) = -\infty\) for all \(y\). Furthermore, if PQP has an optimum of objective value \(v_p\), and if \(W = \{w :Aw \geq b\}\), show that the optimum value of \(LD\) is also \(v_p\). What does this suggest regarding the formulation of \(LD\) for nonconvex situations?

d. Illustrate Part c using the problem to minimize \(\{x_1x_2 : x_1 \geq 0\text{ and } x_2 \geq 0\}\). (This exercise is based on Sherali [1993].)

[6.45] Consider the dual quadratic program given by (6.30). Describe a gradient-based maximization scheme for this problem, following Exercise 6.9. Can you anticipate any computational difficulties? (\textit{Hint:} See Chapter 8.) Illustrate by using the following quadratic programming problem:

\[
\begin{align*}
\text{Minimize} & \quad 3x_1^2 + 2x_2^2 - 2x_1x_2 - 3x_1 - 4x_2 \\
\text{subject to} & \quad 2x_1 + 3x_2 \leq 6 \\
& \quad -x_1 + 2x_2 \leq 2 \\
& \quad x_1, x_2 \geq 0.
\end{align*}
\]

At each iteration, identify the corresponding primal infeasible point as well as the primal feasible point. Develop a suitable measure of infeasibility and check its progress. Can you draw any general conclusions?

[6.46] Let \(X\) and \(Y\) be nonempty sets in \(R^n\), and let \(f, g: R^n \to R\). Consider the \textit{conjugate functions} \(f^*\) and \(g^*\) defined as follows:

\[
\begin{align*}
f^*(u) &= \inf\{f(x) - u'x : x \in X\} \\
g^*(u) &= \sup\{g(x) - u'x : x \in Y\}.
\end{align*}
\]

a. Interpret \(f^*\) and \(g^*\) geometrically.
b. Show that $f^*$ is concave over $X^*$ and $g^*$ is convex over $Y^*$, where $X^* = \{u : f^*(u) > -\infty\}$ and $Y^* = \{u : g^*(u) < \infty\}$.

c. Prove the following conjugate weak duality theorem:
\[
\inf \{f(x) - g(x) : x \in X \cap Y\} \geq \sup \{f^*(u) - g^*(u) : u \in X^* \cap Y^*\}.
\]

d. Now suppose that $f$ is convex, $g$ is concave, $\text{int} X \cap \text{int} Y \neq \emptyset$, and \(\inf \{f(x) - g(x) : x \in X \cap Y\}\) is finite. Show that equality in Part c above holds true and that \(\sup \{f^*(u) - g^*(u) : u \in X^* \cap Y^*\}\) is achieved.

e. By suitable choices of $f$, $g$, $X$, and $Y$, formulate a nonlinear programming problem as follows:
\[
\text{Minimize } f(x) - g(x) \\
\text{subject to } x \in X \cap Y.
\]

What is the form of the conjugate dual problem? Devise some strategies for solving the dual problem.

Notes and References

The powerful results of duality in linear programming and the saddle point optimality criteria for convex programming sparked a great deal of interest in duality in nonlinear programming. Early results in this area include the work of Cottle [1963b], Dorn [1960a], Hanson [1961], Mangasarian [1962], Stoer [1963], and Wolfe [1961].

More recently, several duality formulations that enjoy many of the properties of linear dual programs have evolved. These include the Lagrangian dual problem, the conjugate dual problem, the surrogate dual problem, and the mixed Lagrangian and surrogate, or composite dual, problem. In this chapter we concentrated on the Lagrangian dual formulation because, in our judgment, it is the most promising formulation from a computational standpoint and also because the results of this chapter give the general flavor of the results that one would obtain using other duality formulations. Those interested in studying the subject of conjugate duality may refer to Fenchel [1949], Rockafellar [1964, 1966, 1968, 1969, 1970], Scott and Jefferson [1984, 1989], and Whinston [1967]. For the subject of surrogate duality, where the constraints are grouped into a single constraint by the use of Lagrangian multipliers, refer to Greenberg and Pierskalla [1970b]. Several authors have developed duality formulations that retain the symmetry between the primal and dual problems. The works of Cottle [1963b], Dantzig et al. [1965], Mangasarian and Ponstein [1965], and Stoer [1963] are in this class. For composite duality, see Karwan and Rardin [1979, 1980].

The reader will find the work of Geoffrion [1971b] and Karamardian [1967] excellent references on various duality formulations and their interrelationships. See Everett [1963], Falk [1967, 1969], and Lasdon [1968] for a further study on duality. The relationship between the Lagrangian duality
formulation and other duality formulations is examined in Bazaraa et al. [1971b], Magnanti [1974], and Whinston [1967]. The economic interpretation of duality is covered by Balinski and Baumol [1968], Beckmann and Kapur [1972], Peterson [1970], and Williams [1970].

In Sections 6.1 and 6.2 the dual problem is presented and some of its properties are developed. As a by-product of the main duality theorem, we develop the saddle point optimality criteria for convex programs. These criteria were first developed by Kuhn and Tucker [1951]. For the related concept of min–max duality, see Mangasarian and Ponstein [1965], Ponstein [1965], Rockafellar [1968], and Stoe [1963]. For further discussions and illustrations of perturbation functions, see Geoffrion [1971b] and Minoux [1986]. Larsson and Patriksson [2003] provide a generalized set of near-saddle point optimality conditions and lay the foundation for Lagrangian-based heuristics. For some fundamental discussions and applications of Lagrangian relaxation/dual-based approaches for discrete problems, see Fisher [1981, 1985], Geoffrion [1974], and Shapiro [1979b]. Guignard and Kim [1987] discuss a useful concept of Lagrangian decomposition for exploiting special structures and formulating suitable Lagrangian duals for discrete and nonconvex problems, and Guignard [1998] discusses the value of adding additional constraints (cuts) in a Lagrangian relaxation framework that would have the potential for tightening the relaxation-based bound.

In Section 6.3 we examine several properties of the dual function. We characterize the collection of subgradients at any given point, and use that to determine both ascent directions and the steepest ascent direction. We show that the steepest ascent direction is the shortest subgradient. This result is essentially given by Demyanov [1968]. In Section 6.4 we use these properties to suggest several gradient-based or outer-linearization methods for maximizing the dual function. An accelerated version of the cutting plane method that ensures the generation of ascent directions is discussed by Hearn and Lawphongpanich [1989, 1990]. For a further study of this subject, see Bazaraa and Goode [1979], Demyanov [1968, 1971], Fisher et al. [1975], and Lasdon [1970]. For constraint deletion concepts in outer-linearization methods, see Eaves and Zangwill [1971] and Lasdon [1970]. There are other procedures for solving the dual problem. The cutting plane method discussed in Section 6.4 is a row generation procedure. In its dual form, it is precisely the column generation generalized programming method of Wolfe (see Dantzig [1963]). Another procedure is the subgradient optimization method, which is introduced briefly in Exercises 6.37, 6.38, and 6.39 and is discussed in more detail in Chapter 8. See Held et al. [1974] and Polyak [1967] for validation of subgradient optimization. For related work, see Bazaraa and Goode [1977, 1979], Bazaraa and Sherali [1981], Fisher et al. [1975], Held and Karp [1970], and Sherali et al. [2000].

One of the pioneering works for using the Lagrangian formulation to develop computational schemes is credited to Everett [1963]. Under certain conditions he showed how the primal solution could be retrieved. The result and its extensions are given in Section 6.5. For duality in quadratic programming, see Cottle [1963b], Dorn [1960a,b, 1961a], and Sherali [1993].
Part 3

Algorithms and Their Convergence
Chapter 7

The Concept of an Algorithm

In the remainder of the book, we describe many algorithms for solving different classes of nonlinear programming problems. In this chapter we introduce the concept of an algorithm. Algorithms are viewed as point-to-set maps, and the main convergence theorem is proved utilizing the concept of a closed mapping. This theorem is utilized in the remaining chapters to analyze the convergence of several computational schemes.

Following is an outline of the chapter.

Section 7.1: Algorithms and Algorithmic Maps
In this section we present algorithms as point-to-set maps and introduce the concept of a solution set.

Section 7.2: Closed Maps and Convergence
We introduce the concept of a closed map and prove the main convergence theorem.

Section 7.3: Composition of Mappings
We establish closedness of composite maps by examining closedness of individual maps. We then discuss mixed algorithms and give a condition for their convergence.

Section 7.4: Comparison Among Algorithms
Some practical factors for assessing the efficiency of different algorithms are discussed.

7.1 Algorithms and Algorithmic Maps

Consider the problem to minimize $f(x)$ subject to $x \in S$, where $f$ is the objective function and $S$ is the feasible region. A solution procedure, or an algorithm, for solving this problem can be viewed as an iterative process that generates a sequence of points according to a prescribed set of instructions, together with a termination criterion.

Algorithmic Maps

Given a vector $x_k$ and applying the instructions of the algorithm, we obtain a new point $x_{k+1}$. This process can be described by an algorithmic map $A$. This map is generally a point-to-set map and assigns to each point in the domain $X$ a subset of $X$. Thus, given the initial point $x_1$, the algorithmic map generates the sequence $x_1, x_2, ..., x_{k+1} \in A(x_k)$ for each $k$. The transformation of $x_k$ into $x_{k+1}$ through the map constitutes an iteration of the algorithm.
7.1.1 Example

Consider the following problem:

\[
\text{Minimize } x^2 \\
\text{subject to } x \geq 1,
\]

whose optimal solution is \( x = 1 \). Let the point-to-point algorithmic map be given by \( A(x) = (1/2)(x + 1) \). It can easily be verified that the sequence obtained by applying the map \( A \), with any starting point, converges to the optimal solution \( x = 1 \). With \( x_1 = 4 \), the algorithm generates the sequence \{4, 2.5, 1.75, 1.375, 1.1875, \ldots \}, as illustrated in Figure 7.1a.

As another example, consider the point-to-set mapping \( A \), defined by

\[
A(x) = \begin{cases} 
[l, (1/2)(x + 1)] & \text{if } x \geq 1 \\
[(1/2)(x + 1), 1] & \text{if } x < 1.
\end{cases}
\]

As shown in Figure 7.1b, the image of any point \( x \) is a closed interval, and any point in that interval could be chosen as the successor of \( x \). Starting with any point \( x_1 \), the algorithm converges to \( x = 1 \). With \( x_1 = 4 \), the sequence \{4, 2, 1.2, 1.1, 1.02, \ldots \} is a possible result of the algorithm. Unlike the previous example, other sequences could result from this algorithmic map.

**Solution Set and Convergence of Algorithms**

Consider the following nonlinear programming problem:

\[
\text{Minimize } f(x) \\
\text{subject to } x \in S.
\]
A desirable property of an algorithm for solving the above problem is that it generates a sequence of points converging to a global optimal solution. In many cases, however, we may have to be satisfied with less favorable outcomes. In fact, as a result of nonconvexity, problem size, and other difficulties, we may stop the iterative procedure if a point belonging to a prescribed set, which we call the solution set $\Omega$, is reached. The following are some typical solution sets for the foregoing problem:

1. $\Omega = \{ \bar{x} : \bar{x} \text{ is a local optimal solution of the problem} \}$.
2. $\Omega = \{ \bar{x} : \bar{x} \in S, \ f(\bar{x}) \leq b \}$, where $b$ is an acceptable objective value.
3. $\Omega = \{ \bar{x} : \bar{x} \in S, \ f(\bar{x}) < LB + \varepsilon \}$, where $\varepsilon > 0$ is a specified tolerance and $LB$ is a lower bound on the optimal objective value. A typical lower bound is the objective value of the Lagrangian dual problem.
4. $\Omega = \{ \bar{x} : \bar{x} \in S, \ f(\bar{x}) - v^* < \varepsilon \}$, where $v^*$ is the known global minimum value and $\varepsilon > 0$ is specified.
5. $\Omega = \{ \bar{x} : \bar{x} \text{ satisfies the KKT optimality conditions} \}$.
6. $\Omega = \{ \bar{x} : \bar{x} \text{ satisfies the Fritz John optimality conditions} \}$.

Thus, in general, convergence of algorithms is made in reference to the solution set rather than to the collection of global optimal solutions. In particular, the algorithmic map $A : X \to X$ is said to converge over $Y \subseteq X$ if, starting with any initial point $x_1 \in Y$, the limit of any convergent subsequence of the sequence $x_1, x_2, \ldots$ generated by the algorithm belongs to the solution set $\Omega$. Letting $\Omega$ be the set of global optimal solutions in Example 7.1.1, it is obvious that the two stated algorithms are convergent over the real line with respect to this solution set.

### 7.2 Closed Maps and Convergence

In this section we introduce the notion of closed maps and then prove a convergence theorem. The significance of the concept of closedness will be clear from the following example and the subsequent discussion.

#### 7.2.1 Example

Consider the following problem:

Minimize $x^2$
subject to $x \geq 1$.

Let $\Omega$ be the set of global optimal solutions; that is, $\Omega = \{1\}$. Consider the algorithmic map defined by

$$A(x) = \begin{cases} [3/2 + (1/4)x, 1 + (1/2)x] & \text{if } x \geq 2 \\ (1/2)(x + 1) & \text{if } x < 2. \end{cases}$$
The map $A$ is illustrated in Figure 7.2. Obviously, for any initial point $x_1 \geq 2$, any sequence generated by the map $A$ converges to the point $\hat{x} = 2$. Note that $\hat{x} \notin \Omega$. On the other hand, for $x_1 < 2$, any sequence generated by the algorithm converges to $\overline{x} = 1$. In this example the algorithm converges over the interval $(-\infty, 2)$ but does not converge to a point in the set $\Omega$ over the interval $[2, \infty)$.

Example 7.2.1 shows the significance of the initial point $x_1$, where convergence to a point in $\Omega$ is achieved if $x_1 < 2$ but not realized otherwise. Note that each of the algorithms in Examples 7.1.1 and 7.2.1 satisfies the following conditions:

1. Given a feasible point $x_k \geq 1$, any successor point $x_{k+1}$ is also feasible; that is, $x_{k+1} \geq 1$.
2. Given a feasible point $x_k$ not in the solution set $\Omega$, any successor point $x_{k+1}$ satisfies $f(x_{k+1}) < f(x_k)$, where $f(x) = x^2$. In other words, the objective function strictly decreases.
3. Given a feasible point $x_k$ in the solution set $\Omega$ (i.e., $x_k = 1$), the successor point is also in $\Omega$ (i.e., $x_{k+1} = 1$).

Despite the above-mentioned similarities among the algorithms, the two algorithms of Example 7.1.1 converge to $\overline{x} = 1$, whereas that of Example 7.2.1 does not converge to $\overline{x} = 1$ for any initial point $x_1 \geq 2$. The reason for this is that the algorithmic map of Example 7.2.1 is not closed at $x = 2$. The notion of a
closed mapping, which generalizes the notion of a continuous function, is defined below.

Closed Maps

7.2.2 Definition

Let $X$ and $Y$ be nonempty closed sets in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Let $A: X \rightarrow Y$ be a point-to-set map. The map $A$ is said to be closed at $x \in X$ if for any sequences $\{x_k\}$ and $\{y_k\}$ satisfying

$$
\begin{align*}
x_k &\in X, \\
x_k &\rightarrow x \\
y_k &\in A(x_k), \\
y_k &\rightarrow y
\end{align*}
$$

we have that $y \in A(x)$. The map $A$ is said to be closed on $Z \subseteq X$ if it is closed at each point in $Z$.

Figure 7.2 shows an example of a point-to-set map that is not closed at $x = 2$. In particular, the sequence $\{x_k\}$ with $x_k = 2 - 1/k$ converges to $x = 2$, and the sequence $\{y_k\}$ with $y_k = A(x_k) = 3/2 - 1/(2k)$ converges to $y = 3/2$, but $y \not\in A(x) = \{2\}$. Figure 7.1 shows two examples of algorithmic maps that are closed everywhere.

Zangwill’s Convergence Theorem

Conditions that ensure convergence of algorithmic maps are stated in Theorem 7.2.3, which is due to Willard Zangwill. The theorem is used in the remainder of the book to show convergence of many algorithms.

7.2.3 Theorem

Let $X$ be a nonempty closed set in $\mathbb{R}^n$, and let the nonempty set $\Omega \subseteq X$ be the solution set. Let $A: X \rightarrow X$ be a point-to-set map. Given $x_1 \in X$, the sequence $\{x_k\}$ is generated iteratively as follows: If $x_k \in \Omega$, then stop; otherwise, let $x_{k+1} \in A(x_k)$, replace $k$ by $k + 1$, and repeat.

Suppose that the sequence $x_1, x_2, \ldots$ produced by the algorithm is contained in a compact subset of $X$, and suppose that there exists a continuous function $\alpha$, called the descent function, such that $\alpha(y) < \alpha(x)$ if $x \not\in \Omega$ and $y \in A(x)$. If the map $A$ is closed over the complement of $\Omega$, then either the algorithm stops in a finite number of steps with a point in $\Omega$ or it generates an infinite sequence $\{x_k\}$ such that:

1. Every convergent subsequence of $\{x_k\}$ has a limit in $\Omega$; that is, all accumulation points of $\{x_k\}$ belong to $\Omega$. 
2. \( \alpha(x_k) \to \alpha(x) \) for some \( x \in \Omega \).

**Proof**

If at any iteration a point \( x_k \) in \( \Omega \) is generated, then the algorithm stops. Now suppose that an infinite sequence \( \{x_k\} \) is generated. Let \( \{x_k\} \) be any convergent subsequence with limit \( x \in X \). Since \( \alpha \) is continuous, then for \( k \in \mathcal{K} \)
\[ \alpha(x_k) \to \alpha(x). \]

Thus, for a given \( \epsilon > 0 \), there is a \( K \in \mathcal{K} \) such that

\[ \alpha(x_k) - \alpha(x) < \epsilon \quad \text{for } k \geq K \text{ with } k \in \mathcal{K} \]

In particular, for \( k = K \) we get

\[ \alpha(x_k) - \alpha(x) < \epsilon. \quad (7.1) \]

Now let \( k > K \). Since \( \alpha \) is a descent function, \( \alpha(x_k) < \alpha(x_K) \), and from (7.1), we get

\[ \alpha(x_k) - \alpha(x) = \alpha(x_k) - \alpha(x_K) + \alpha(x_K) - \alpha(x) < 0 + \epsilon = \epsilon. \]

Since this is true for every \( k > K \), and since \( \epsilon > 0 \) was arbitrary,

\[ \lim_{k \to \infty} \alpha(x_k) = \alpha(x). \quad (7.2) \]

We now show that \( x \in \Omega \). By contradiction, suppose that \( x \notin \Omega \), and consider the sequence \( \{x_{k+1}\} \). This sequence is contained in a compact subset of \( X \) and hence has a convergent subsequence \( \{x_{k+1}\} \) with limit \( \bar{x} \) in \( X \). Noting (7.2), it is clear that \( \alpha(\bar{x}) = \alpha(x) \). Since \( A \) is closed at \( x \), and for \( k \in \mathcal{K} \), \( x_k \to x \), \( x_{k+1} \in A(x_k) \), and \( x_{k+1} \to \bar{x} \), then \( \bar{x} \in A(x) \). Therefore, \( \alpha(\bar{x}) \) is a contradiction to the fact that \( \alpha(\bar{x}) = \alpha(x) \). Thus, \( x \in \Omega \) and Part 1 of the theorem holds true. This, coupled with (7.2), shows that Part 2 of the theorem holds true, and the proof is complete.

**Corollary**

Under the assumptions of the theorem, if \( \Omega \) is the singleton \( \{\bar{x}\} \), then the entire sequence \( \{x_k\} \) converges to \( \bar{x} \).

**Proof**

Suppose, by contradiction, that there exist an \( \epsilon > 0 \) and a sequence \( \{x_k\} \) such that

\[ \|x_k - \bar{x}\| > \epsilon \quad \text{for } k \in \mathcal{K} \quad (7.3) \]
Note that there exists \( \mathcal{H}' \subset \mathcal{H} \) such that \( \{x_k\}_{\mathcal{H}'} \) has a limit \( x' \). By Part 1 of the theorem, \( x' \in \Omega \). But \( \Omega = \{\bar{x}\} \), and thus \( x' = \bar{x} \). Therefore, \( x_k \to \bar{x} \) for \( \mathcal{H} \in \mathcal{H}' \), violating (7.3). This completes the proof.

Note that if the point at hand \( x_k \) does not belong to the solution set \( \Omega \), the algorithm generates a new point \( x_{k+1} \) such that \( \alpha(x_{k+1}) < \alpha(x_k) \). As mentioned before, the function \( \alpha \) is called a descent function. In many cases, \( \alpha \) is chosen as the objective function \( f \) itself, and thus the algorithm generates a sequence of points with improving objective function values. Other alternative choices of the function \( \alpha \) are possible. For instance, if \( f \) is differentiable, \( \alpha \) could be chosen as \( \alpha(x) = \|\nabla f(x)\| \) for an unconstrained optimization problem, since we know that \( \nabla f(\bar{x}) = 0 \) for any (local/global) optimum \( \bar{x} \).

**Terminating the Algorithm**

As indicated in Theorem 7.2.3, the algorithm is terminated if we reach a point in the solution set \( \Omega \). In most cases, however, convergence to a point in \( \Omega \) occurs only in a limiting sense, and we must resort to some practical rules for terminating the iterative procedure. The following rules are frequently used to stop a given algorithm. Here \( \varepsilon > 0 \) and the positive integer \( N \) are prespecified.

1. \( \|x_{k+N} - x_k\| < \varepsilon \).
   Here, the algorithm is stopped if the distance moved after \( N \) applications of the map \( A \) is less than \( \varepsilon \).

2. \( \|x_{k+1} - x_k\| < \varepsilon \).
   \( \|x_k\| < \varepsilon \).
   Under this criterion, the algorithm is terminated if the relative distance moved during a given iteration is less than \( \varepsilon \).

3. \( \alpha(x_k) - \alpha(x_{k+N}) < \varepsilon \).
   Here, the algorithm is stopped if the total improvement in the descent function value after \( N \) applications of the map \( A \) is less than \( \varepsilon \).

4. \( \frac{\alpha(x_k) - \alpha(x_{k+1})}{\alpha(x_k)} < \varepsilon \).
   If the relative improvement in the descent function value during any given iteration is less than \( \varepsilon \), then this termination criterion is realized.

5. \( \alpha(x_k) - \alpha(\bar{x}) < \varepsilon \), where \( \bar{x} \) belongs to \( \Omega \).
   This criterion for termination is suitable if \( \alpha(\bar{x}) \) is known beforehand; for example, in unconstrained optimization, if \( \alpha(x) = \|\nabla f(x)\| \) and \( \Omega = \{\bar{x} : \nabla f(\bar{x}) = 0\} \), then \( \alpha(\bar{x}) = 0 \).
7.3 Composition of Mappings

In most nonlinear programming solution procedures, the algorithmic maps are often composed of several maps. For example, some algorithms first find a direction $d_k$ to move along and then determine the step size $\lambda_k$ by solving the one-dimensional problem of minimizing $a(x_k + \lambda d_k)$. In this case, the map $A$ is composed of $MD$, where $D$ finds the direction $d_k$ and then $M$ finds an optimal step size $\lambda_k$. It is often easier to prove that the overall map is closed by examining its individual components. In this section the notion of composite maps is stated precisely, and then a result relating closedness of the overall map to that of its individual components is given. Finally, we discuss mixed algorithms and state conditions under which they converge.

7.3.1 Definition

Let $X$, $Y$, and $Z$ be nonempty closed sets in $\mathbb{R}^n$, $\mathbb{R}^p$, and $\mathbb{R}^q$, respectively. Let $B: X \rightarrow Y$ and $C: Y \rightarrow Z$ be point-to-set maps. The composite map $A = CB$ is defined as the point-to-set map $A: X \rightarrow Z$ with

$$A(x) = \bigcup \{C(y) : y \in B(x)\}.$$ 

Figure 7.3 illustrates the notion of a composite map, and Theorem 7.3.2 and its corollaries give several sufficient conditions for a composite map to be closed.

7.3.2 Theorem

Let $X$, $Y$, and $Z$ be nonempty closed sets in $\mathbb{R}^n$, $\mathbb{R}^p$, and $\mathbb{R}^q$, respectively. Let $B: X \rightarrow Y$ and $C: Y \rightarrow Z$ be point-to-set maps, and consider the composite map $A = CB$. Suppose that $B$ is closed at $x$ and that $C$ is closed on $B(x)$. Furthermore, suppose that if $x_k \rightarrow x$ and $y_k \in B(x_k)$, then there is a convergent subsequence of $\{y_k\}$. Then $A$ is closed at $x$. 

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{composite_maps.png}
\caption{Composite maps.}
\end{figure}
Proof

Let \( x_k \to x \), \( z_k \in A(x_k) \), and \( z_k \to z \). We need to show that \( z \in A(x) \).

By the definition of \( A \), for each \( k \) there is a \( y_k \in B(x_k) \) such that \( z_k \in C(y_k) \). By assumption, there is a convergent subsequence \( \{ y_k \} \) with limit \( y \). Since \( B \) is closed at \( x \), then \( y \in B(x) \). Furthermore, since \( C \) is closed on \( B(x) \), it is closed at \( y \), and hence \( z \in C(y) \). Thus, \( z \in C(y) \subset CB(x) = A(x) \), and hence \( A \) is closed at \( x \).

Corollary 1

Let \( X \), \( Y \), and \( Z \) be nonempty closed sets in \( R^n \), \( R^p \), and \( R^q \), respectively. Let \( B : X \to Y \) and \( C : Y \to Z \) be point-to-set maps. Suppose that \( B \) is closed at \( x \), \( C \) is closed on \( B(x) \), and \( Y \) is compact. Then \( A = CB \) is closed at \( x \).

Corollary 2

Let \( X \), \( Y \), and \( Z \) be nonempty closed sets in \( R^n \), \( R^p \), and \( R^q \), respectively. Let \( B : X \to Y \) be a function, and let \( C : Y \to Z \) be a point-to-set map. If \( B \) is continuous at \( x \), and \( C \) is closed on \( B(x) \), then \( A = CB \) is closed at \( x \).

Note the importance of the assumption that a convergent subsequence \( \{ y_k \} \) exists in Theorem 7.3.2. Without this assumption, even if the maps \( B \) and \( C \) are closed, the composite map \( A = CB \) is not necessarily closed, as shown by Example 7.3.3 (due to Jamie J. Goode).

7.3.3 Example

Consider the maps \( B, C : R \to R \) defined as

\[
B(x) = \begin{cases} 
1/x & \text{if } x \neq 0 \\
0 & \text{if } x = 0
\end{cases}
\]

\[
C(y) = \{ z : |z| \leq |y| \}.
\]

Note that both \( B \) and \( C \) are closed everywhere. (Observe that the closedness of \( B \) at \( x = 0 \) holds true vacuously because for \( \{ x_k \} \to 0^+ \), the corresponding sequence \( \{ y_k \} = \{ B(x_k) \} \) does not have a limit point.) Now consider the composite map \( A = CB \). Then \( A \) is given by \( A(x) = CB(x) = \{ z : |z| \leq |B(x)| \} \).

From the definition of \( B \) it follows that

\[
A(x) = \begin{cases} 
\{ z : |z| \leq \frac{1}{|x|} \} & \text{if } x \neq 0 \\
\{ 0 \} & \text{if } x = 0.
\end{cases}
\]

Note that \( A \) is not closed at \( x = 0 \). In particular, consider the sequence \( \{ x_k \} \), where \( x_k = 1/k \). Note that \( A(x_k) = \{ z : |z| \leq k \} \), and hence \( z_k = 1 \) belongs to
A(x_k) for each k. On the other hand, the limit point \( z = 1 \) does not belong to \( A(x) = \{0\} \). Thus, the map \( A \) is not closed, even though both \( B \) and \( C \) are closed. Here, Theorem 7.3.2 does not apply, since the sequence \( y_k \in B(x_k) \) for \( x_k = 1/k \) does not have a convergent subsequence.

**Convergence of Algorithms with Composite Maps**

At each iteration, many nonlinear programming algorithms use two maps, \( B \) and \( C \), say. One of the maps, \( B \), is usually closed and satisfies the convergence requirements of Theorem 7.2.3. The second map, \( C \), may involve any process as long as the value of the descent function does not increase. As illustrated in Exercise 7.1, the overall map may not be closed, so that Theorem 7.2.3 cannot be applied. However, as shown below, such maps do converge. Hence, such a result can be used to establish the convergence of a complex algorithm in which a step of a known convergent algorithm is interspersed at finite iteration intervals, but infinitely often over the entire algorithmic sequence. Then, by viewing the algorithm as an application of the composite map \( CB \), where \( B \) corresponds to the step of the known convergent algorithm that satisfies the assumptions of Theorem 7.2.3 and \( C \) corresponds to the set of intermediate steps of the complex algorithm, the overall convergence of such a scheme would follow by Theorem 7.3.4. In such a context, the step of applying \( B \) as above is called a *spacer step*.

**7.3.4 Theorem**

Let \( X \) be a nonempty closed set in \( \mathbb{R}^n \), and let \( \Omega \subseteq X \) be a nonempty solution set. Let \( \alpha : \mathbb{R}^n \to \mathbb{R} \) be a continuous function, and consider the point-to-set map \( C : X \to X \) satisfying the following property: Given \( x \in X \), then \( \alpha(y) \leq \alpha(x) \) for \( y \in C(x) \). Let \( B : X \to X \) be a point-to-set map that is closed over the complement of \( \Omega \) and that satisfies \( \alpha(y) < \alpha(x) \) for each \( y \in B(x) \) if \( x \notin \Omega \). Now consider the algorithm defined by the composite map \( A = CB \). Given \( x_1 \in X \), suppose that the sequence \( \{x_k\} \) is generated as follows: If \( x_k \in \Omega \), then stop; otherwise, let \( x_{k+1} \in A(x_k) \), replace \( k \) by \( k + 1 \), and repeat. Suppose that the set \( \Lambda = \{x : \alpha(x) \leq \alpha(x_1)\} \) is compact. Then either the algorithm stops in a finite number of steps with a point in \( \Omega \) or all accumulation points of \( \{x_k\} \) belong to \( \Omega \).

**Proof**

If at any iteration \( x_k \in \Omega \), the algorithm stops finitely. Now, suppose that the sequence \( \{x_k\} \) is generated by the algorithm, and let \( \{x_k\} \to x \) be a convergent subsequence, with limit \( x \). Thus, \( \alpha(x_k) \to \alpha(x) \) for \( k \in \mathbb{R} \). Using the monotonicity of \( \alpha \) as in Theorem 7.2.3, it follows that

\[
\lim_{k \to +\infty} \alpha(x_k) = \alpha(x). \tag{7.4}
\]
We want to show that \( x \in \Omega \). By contradiction, suppose that \( x \notin \Omega \), and consider the sequence \( \{x_{k+1}\}_{k \in \mathbb{N}} \). By the definition of the composite map \( \Lambda \), note that \( x_{k+1} \in C(y_k) \), where \( y_k \in B(x_k) \). Note that \( y_k, x_{k+1} \in \Lambda \). Since \( \Lambda \) is compact, there exists an index set \( \mathcal{K} \subseteq \mathbb{N} \) such that \( y_k \to y \) and \( x_{k+1} \to x' \) for \( k \in \mathcal{K} \). Since \( B \) is closed at \( x \in \Omega \), then \( y \in B(x) \), and \( \alpha(y) < \alpha(x) \). Since \( x_{k+1} \in C(y_k) \), then, by assumption, \( \alpha(x_{k+1}) \leq \alpha(y_k) \) for \( k \in \mathcal{K} \); and hence by taking the limit, \( \alpha(x') \leq \alpha(y) \). Since \( \alpha(y) < \alpha(x) \), then \( \alpha(x') < \alpha(x) \). Since \( \alpha(x_{k+1}) \to \alpha(x') \) for \( k \in \mathcal{K} \), then \( \alpha(x') < \alpha(x) \) contradicts (7.4). Therefore, \( x \in \Omega \), and the proof is complete.

**Minimizing Along Independent Directions**

We now present a theorem that establishes convergence of a class of algorithms for solving a problem of the form: Minimize \( f(x) \) subject to \( x \in \mathbb{R}^n \). Under mild assumptions, we show that an algorithm that generates \( n \) linearly independent search directions, and obtains a new point by sequentially minimizing \( f \) along these directions, converges to a stationary point. The theorem also establishes convergence of algorithms using linearly independent and orthogonal search directions.

**7.3.5 Theorem**

Let \( f: \mathbb{R}^n \to \mathbb{R} \) be differentiable, and consider the problem to minimize \( f(x) \) subject to \( x \in \mathbb{R}^n \). Consider an algorithm whose map \( A \) is defined as follows. The vector \( y \in A(x) \) means that \( y \) is obtained by minimizing \( f \) sequentially along the directions \( d_1, \ldots, d_n \) starting from \( x \). Here, the search directions \( d_1, \ldots, d_n \) may depend on \( x \), and each has norm 1. Suppose that the following properties are true:

1. There exists an \( \varepsilon > 0 \) such that \( \det[D(x)] \geq \varepsilon \) for each \( x \in \mathbb{R}^n \). Here \( D(x) \) is the \( n \times n \) matrix whose columns are the search directions generated by the algorithm, and \( \det[D(x)] \) denotes the determinant of \( D(x) \).

2. The minimum of \( f \) along any line in \( \mathbb{R}^n \) is unique.

Given a starting point \( x_1 \), suppose that the algorithm generates the sequence \( \{x_k\} \) as follows. If \( \nabla f(x_k) = 0 \), then the algorithm stops with \( x_k \); otherwise, \( x_{k+1} \in A(x_k) \), \( k \) is replaced by \( k + 1 \), and the process is repeated. If the sequence \( \{x_k\} \) is contained in a compact subset of \( \mathbb{R}^n \), then each accumulation point \( x \) of the sequence \( \{x_k\} \) must satisfy \( \nabla f(x) = 0 \).
Proof

If the sequence \( \{x_k\} \) is finite, then the result is immediate. Now suppose that the algorithm generates the infinite sequence \( \{x_k\} \).

Let \( \mathcal{H} \) be an infinite sequence of positive integers, and suppose that the sequence \( \{x_k\}_{\mathcal{H}} \) converges to a point \( x \). We need to show that \( \forall f(x) = 0 \).

Suppose by contradiction that \( \forall f(x) \neq 0 \), and consider the sequence \( \{x_{k+1}\}_{\mathcal{H}} \).

By assumption, this sequence is contained in a compact subset of \( \mathbb{R}^n \); and hence there exists \( \mathcal{H}' \subseteq \mathcal{H} \) such that \( \{x_{k+1}\}_{\mathcal{H}'} \) converges to \( x' \). We show first that \( x' \) can be obtained from \( x \) by minimizing \( f \) along a set of \( n \) linearly independent directions.

Let \( D_k \) be the \( n \times n \) matrix whose columns \( d_{1k}, \ldots, d_{nk} \) are the search directions generated at iteration \( k \). Thus, \( x_{k+1} = x_k + D_k \lambda_k = x_k + \sum_{j=1}^n d_{jk} \lambda_{jk} \), where \( \lambda_{jk} \) is the distance moved along \( d_{jk} \). In particular, letting \( y_{jk} = x_k \), \( y_{j+1, k} = y_j + \lambda_{jk} d_{jk} \) for \( j = 1, \ldots, n \), it follows that \( x_{k+1} = y_{n+1, k} \) and

\[
\forall \lambda \in \mathbb{R} \quad f(y_{j+1, k}) \leq f(y_{jk} + \lambda d_{jk}) \quad \text{for all } j = 1, \ldots, n. \tag{7.5}
\]

Since \( \det[D_k] \geq \varepsilon > 0 \), \( D_k \) is invertible, so that \( \lambda_k = D_k^{-1}(x_{k+1} - x_k) \). Since each column of \( D_k \) has norm 1, there exists \( \mathcal{H}^* \subseteq \mathcal{H} \) such that \( D_k \rightarrow D \). Since \( \det[D_k] \geq \varepsilon \) for each \( k \), \( \det[D] \geq \varepsilon \), so that \( D \) is invertible. Now for \( k \in \mathcal{H}^* \), \( x_{k+1} \rightarrow x' \), \( x_k \rightarrow x \), \( D_k \rightarrow D \), so that \( \lambda_k \rightarrow \lambda \), where \( \lambda = D^{-1}(x' - x) \).

Therefore, \( x' = x + D \lambda = x + \sum_{j=1}^n d_{j} \lambda_j \). Let \( y_1 = x \), and for \( j = 1, \ldots, n \), let \( y_{j+1} = y_j + \lambda_j d_j \), so that \( x' = y_{n+1} \). To show that \( x' \) is obtained from \( x \) by minimizing \( f \) sequentially along the directions \( d_1, \ldots, d_n \), it suffices to show that

\[
\forall \lambda \in \mathbb{R} \quad f(y_{j+1}) \leq f(y_{j} + \lambda d_{j}) \quad \text{for all } j = 1, \ldots, n. \tag{7.6}
\]

Note that \( \lambda_{jk} \rightarrow \lambda_j \), \( d_{jk} \rightarrow d_j \), \( x_k \rightarrow x \), and \( x_{k+1} \rightarrow x' \) as \( k \in \mathcal{H}^* \) approaches \( \infty \), so that \( y_{jk} \rightarrow y_j \) for \( j = 1, \ldots, n+1 \) as \( k \in \mathcal{H}^* \) approaches \( \infty \). By the continuity of \( f \), (7.6) follows from (7.5). We have thus shown that \( x' \) is obtained from \( x \) by minimizing \( f \) sequentially along the directions \( d_1, \ldots, d_n \).

Obviously, \( f(x') \leq f(x) \). First, consider the case \( f(x') < f(x) \). Since \( \{f(x_k)\} \) is a nonincreasing sequence, and since \( f(x_k) \rightarrow f(x) \) as \( k \in \mathcal{H} \) approaches \( \infty \), \( \lim_{k \rightarrow \infty} f(x_k) = f(x) \). This is impossible, however, in view of the fact that \( x_{k+1} \rightarrow x' \) as \( k \in \mathcal{H}' \) approaches \( \infty \) and the assumption that \( f(x') < f(x) \). Now, consider the case \( f(x') = f(x) \). By Property 2 of the theorem, and since \( x' \) is obtained from \( x \) by minimizing \( f \) along \( d_1, \ldots, d_n \), \( x' = x \). This implies
The Concept of an Algorithm

further that $\nabla f(x)'d_j = 0$ for $j = 1, \ldots, n$. Since $d_1, \ldots, d_n$ are linearly independent, we get that $\nabla f(x) = 0$, contradicting our assumption. This completes the proof.

Note that no closedness or continuity assumptions are made on the map that provides the search directions. It is only required that the search directions used at each iteration be linearly independent and that as these directions converge, the limiting directions must also be linearly independent. Obviously this holds true if a fixed set of linearly independent search directions are used at every iteration. Alternatively, if the search directions used at each iteration are mutually orthogonal and each has norm 1, then the search matrix $D$ satisfies $D'D = I$. Therefore, $\det[D] = 1$, so that Condition 1 of the theorem holds true.

Also, note that Condition 2 in the statement of the theorem is used to ensure the following property. If a differentiable function $f$ is minimized along $n$ independent directions starting from a point $x$ and resulting in $x'$, then $f(x') < f(x)$, provided that $\nabla f(x) \neq 0$. Without Assumption 2, this is not true, as evidenced by $f(x_1, x_2) = x_2(1 - x_1)$. If $x = (0, 0)'$, then minimizing $f$ starting from $x$ along $d_1 = (1, 0)'$ and then along $d_2 = (0, 1)'$ could produce the point $x' = (1, 1)'$, where $f(x') = f(x) = 0$, even though $\nabla f(x) = (0, 1)' \neq (0, 0)'$.

7.4 Comparison Among Algorithms

In the remainder of the book, we discuss several algorithms for solving different classes of nonlinear programming problems. In this section we discuss some important factors that must be considered when assessing the effectiveness of these algorithms and when comparing them. These factors are (1) generality, reliability, and precision; (2) sensitivity to parameters and data; (3) preparational and computational effort; and (4) convergence.

Generality, Reliability, and Precision

Different algorithms are designed for solving various classes of nonlinear programming problems, such as unconstrained optimization problems and problems having inequality constraints, equality constraints, or both types of constraints. Within each of these classes, different algorithms make specific assumptions about the problem structure. For example, for unconstrained optimization problems, some procedures assume that the objective function is differentiable, whereas other algorithms do not make this assumption and rely primarily on functional evaluations only. For problems having equality constraints, some algorithms can handle only linear constraints, whereas others can handle nonlinear constraints as well. Thus, generality of an algorithm refers to the variety of problems that the algorithm can handle and also to the restrictiveness of the assumptions required by the algorithm.

Another important factor is the reliability, or robustness, of the algorithm. Given any algorithm, it is not difficult to construct a test problem that it cannot
solve effectively even if the problem satisfies all the assumptions required. Reliability, or robustness, means the ability of the procedure to solve most of the problems in the class for which it is designed with reasonable accuracy. Usually, this characteristic should hold regardless of the starting (feasible) solution used. The relationship between the reliability of a certain procedure and the problem size and structure cannot be overlooked. Some algorithms are reliable if the number of variables is small or if the constraints are not highly nonlinear, and are not reliable otherwise.

As implied by Theorem 7.2.3, convergence of nonlinear programming algorithms usually occurs in a limiting sense, if at all. Thus, we are interested in measuring the quality of the points produced by the algorithm after a reasonable number of iterations. Algorithms that quickly produce feasible solutions with good objective values are preferred. As discussed in Chapter 6 and as will be seen in Chapter 9, several procedures generate a sequence of infeasible solutions, where feasibility is achieved only at termination. Hence, at later iterations, it is imperative that the degree of infeasibility be small so that a near-feasible solution will be at hand if the algorithmic process is terminated prematurely.

**Sensitivity to Parameters and Data**

For most algorithms, the user must set initial values for certain parameters, such as the starting vector, the step size, the acceleration factor, and parameters for terminating the algorithm. Some procedures are quite sensitive to these parameters and to the problem data and may produce different results or stop prematurely, depending on their values. In particular, for a fixed set of selected parameters, the algorithm should solve the problem for a wide range of problem data and should be *scale invariant*, that is, insensitive to any constraint or variable scaling that might be used. Similarly, for a given set of problem data, one would prefer that the algorithm not be very sensitive to selected values of the parameters. (See Section 1.3 for a related discussion.)

**Preparational and Computational Effort**

Another basis for comparing algorithms is the total effort, both preparational and computational, expended for solving problems. The effort of preparing the input data should be taken into consideration when evaluating an algorithm. An algorithm that uses first- or second-order derivatives, especially if the original functions are complicated, requires a considerably larger amount of preparation time than one that uses only functional evaluations. The computational effort of an algorithm is usually assessed by the computer time, the number of iterations, or the number of functional evaluations. However, any of these measures, by itself, is not entirely satisfactory. The computer time needed to execute an algorithm depends not only on its efficiency but also on the type of machine used, the character of the time measured, the existing load on the machine, and the efficiency of coding. Also, the number of iterations cannot be used as the only measure of effectiveness of an algorithm because the effort per iteration may vary considerably from one procedure to another. Finally, the number of func-
tional evaluations can be misleading, since it does not measure other operations, such as matrix multiplication, matrix inversion (or factorizations; see Appendix A.2), and finding suitable directions of movement. In addition, for derivative-dependent methods, we have to weigh the evaluation of first- and second-order derivatives against the evaluation of the functions themselves and their net consequence on algorithmic performance.

Convergence

Theoretical convergence of algorithms to points in the solution set is a highly desirable property. Given two competing algorithms that converge, they could be compared theoretically on the basis of the order or speed of convergence. This notion is defined below.

7.4.1 Definition

Let the sequence \( \{r_k\} \) of real numbers converge to \( \bar{r} \), and assume that \( r_k \neq \bar{r} \) for all \( k \). The order of convergence of the sequence is the supremum of the non-negative numbers \( p \) satisfying

\[
\lim_{k \to \infty} \frac{|r_{k+1} - \bar{r}|}{|r_k - \bar{r}|^p} = \beta < \infty.
\]

If \( p = 1 \) and the convergence ratio \( \beta \in (0, 1) \), the sequence is said to have a linear convergence rate. Since asymptotically we have \( |r_{k+1} - \bar{r}| = \beta|r_k - \bar{r}| \), linear convergence is sometimes also referred to as geometric convergence, although often, this terminology is reserved for situations in which the sequence is truly a geometric sequence. If \( p > 1 \), or if \( p = 1 \) and \( \beta = 0 \), the sequence is said to have superlinear convergence. In particular, if \( p = 2 \) and \( \beta < \infty \), the sequence is said to have a second-order, or quadratic, rate of convergence.

For example, the sequence \( r_k \) of iterates generated by the algorithmic map of Figure 7.1a satisfies \( r_{k+1} = (r_k + 1)/2 \), where \( \{r_k\} \to 1 \). Hence, \( (r_{k+1} - 1) = (r_k - 1)/2 \); so, with \( p = 1 \), the limit in Definition 7.4.1 is \( \beta = 1/2 \). However, for \( p > 1 \), this limit is infinity. Consequently, \( \{r_k\} \to 1 \) only linearly.

On the other hand, suppose that we have \( r_{k+1} = 1 + (r_k - 1)/2^k \) for \( k = 1, 2, \ldots \), where \( \eta = 4 \), say. In lieu of the sequence \( \{4, 1.75, 1.375, 1.1875, \ldots\} \) obtained above, we now produce the sequence \( \{4, 2.5, 1.375, 1.046875, \ldots\} \). The sequence can readily be verified to be converging to unity. However, we now have \( |r_{k+1} - 1|/|r_k - 1| = 1/2^k \), which approaches 0 as \( k \to \infty \). Hence, \( \{r_k\} \to 1 \) superlinearly in this case.

If \( r_k \) in Definition 7.4.1 represents \( \alpha(x_k) \), the value of the descent function at the \( k \)th iteration, then the larger the value of \( p \), the faster is the convergence of the algorithm. If the limit in Definition 7.4.1 exists, then for large
values of $k$, we have, asymptotically, $|r_{k+1} - \bar{r}| = \beta |r_k - \bar{r}|^p$, which indicates faster convergence for larger values of $p$. For the same value of $p$, the smaller the convergence ratio $\beta$, the faster is the convergence. It should be noted, however, that the order of convergence and the ratio of convergence must not be used solely for evaluating algorithms that converge, since they represent the progress of the algorithm only as the number of iterations approach infinity. (See the Notes and References section for readings on average rates of convergence that deal with the average progress per step achieved over a large number of iterations, in contrast with the stepwise progress discussed above.)

In a similar manner, we can define convergence rates for a vector sequence $\{x_k\} \to \bar{x}$. Again, let us suppose that $x_k \neq \bar{x}$ for all $k$ (or, alternatively, for $k$ large enough). We can now define rates of convergence with respect to an error function that measures the separation between $x_k$ and $\bar{x}$, typically the Euclidean distance function $\|x_k - \bar{x}\|$. Consequently, in Definition 7.4.1 we simply replace $|r_k - \bar{r}|$ by $\|x_k - \bar{x}\|$ for all $k$. In particular, if there exists a $0 < \rho < 1$ such that $\|x_{k+1} - \bar{x}\| \leq \rho \|x_k - \bar{x}\|$ for all $k$, then $\{x_k\}$ converges to $\bar{x}$ at a linear rate. On the other hand, if $\|x_{k+1} - \bar{x}\| \leq \rho_k \|x_k - \bar{x}\|$ for all $k$, where $\{\rho_k\} \to 0$, then the rate of convergence is superlinear. Note that these are only frequently used interpretations that coincide with Definition 7.4.1, using $\|x_k - \bar{x}\|$ in place of $|r_k - \bar{r}|$.

We also point out here that the foregoing rates of convergence, namely, linear, superlinear, quadratic, and so on, are sometimes referred to, respectively, as $q$-linear, $q$-superlinear, $q$-quadratic, and so on. The prefix $q$ stands for the quotient taken in Definition 7.4.1 and is used to differ from another weaker type of $r$-(root)-order convergence rate, in which the errors $\|x_k - \bar{x}\|$ are bounded above only by elements of some $q$-order sequence converging to zero (see the Notes and References section).

Another convergence criterion frequently used in comparing algorithms is their ability to effectively minimize quadratic functions. This is used because, near the minimum, a linear approximation to a function is poor, whereas it can be adequately approximated by a quadratic form. Thus, an algorithm that does not perform well for minimizing a quadratic function is unlikely to perform well for a general nonlinear function as we move closer to optimality.

Exercises

[7.1] This exercise illustrates that a map for a convergent algorithm need not be closed. Consider the following problem:

Minimize $x^2$

subject to $x \in R$.

Consider the maps $B, C : R \to R$ defined as
The Concept of an Algorithm

\[
B(x) = \frac{x}{2} \quad \text{for all } x;
\]

\[
C(x) = \begin{cases} 
    x & \text{if } -1 \leq x \leq 1 \\
    x + 1 & \text{if } x < -1 \\
    x - 1 & \text{if } x > 1.
\end{cases}
\]

Let the solution set \( \Omega = \{0\} \), and let the descent function \( \alpha(x) = x^2 \).

a. Show that \( B \) and \( C \) satisfy all the assumptions of Theorem 7.3.4.

b. Verify that the composite map \( A = CB \) is as given below, and verify that it is not closed:

\[
A(x) = \begin{cases} 
    x/2 & \text{if } -2 \leq x \leq 2 \\
    (x/2) + 1 & \text{if } x < -2 \\
    (x/2) - 1 & \text{if } x > 2.
\end{cases}
\]

c. Despite the fact that \( A \) is not closed, show that the algorithm defined by \( A \) converges to the point \( \bar{x} = 0 \), regardless of the starting point.

[7.2] Which of the following maps are closed and which are not?

a. \( A(x) = \{y : x^2 + y^2 \leq 2\} \).

b. \( A(x) = \{y : x'y \leq 2\} \).

c. \( A(x) = \{y : \|y - x\| \leq 2\} \).

d. \( A(x) = \begin{cases} 
    \{y : x^2 + y^2 \leq 1\} & \text{if } x \neq 0 \\
    [-1, 0] & \text{if } x = 0.
\end{cases} \)

[7.3] Let \( A : \mathbb{R}^n \rightarrow \mathbb{R}^n \) be the point-to-set map defined as follows. Given an \( m \times n \) matrix \( B \), an \( m \)-vector \( b \), and an \( n \)-vector \( x \), then \( y \in A(x) \) means that \( y \) is an optimal solution to the problem to minimize \( x'z \) subject to \( Bz = b, z \geq 0 \). Show that the map \( A \) is closed.

[7.4] Let \( A : \mathbb{R}^m \rightarrow \mathbb{R}^n \) be the point-to-set map defined as follows. Given an \( m \times n \) matrix \( B \), an \( n \)-vector \( c \), and an \( m \)-vector \( x \), then \( y \in A(x) \) means that \( y \) is an optimal solution to the problem to minimize \( c'z \) subject to \( Bz = x, z \geq 0 \).

a. Show that the map \( A \) is closed at \( x \) if the set \( Z = \{z : Bz = x, z \geq 0\} \) is compact.

b. What are your conclusions if the set \( Z \) is not compact?

[7.5] Which of the following maps are closed, and which are not?

a. \((y_1, y_2) \in A(x_1, x_2) \) means that \( y_1 = x_1 - 1 \) and \( y_2 \in [x_2 - 1, x_2 + 1] \).
Let \( X \) and \( Y \) be nonempty closed sets in \( \mathbb{R}^p \) and \( \mathbb{R}^q \), respectively. Let \( A: X \to Y \) and \( B: X \to Y \) be point-to-set maps. The sum map \( C = A + B \) is defined by \( C(x) = \{a + b : a \in A(x), b \in B(x)\} \). Show that if \( A \) and \( B \) are closed and if \( Y \) is compact, \( C \) is closed.

Let \( A: \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n \) be the point-to-set map defined as follows. Given \( x, z \in \mathbb{R}^n \), then \( y \in A(x, z) \) means that \( y = \lambda x + (1 - \lambda)z \) for some \( \lambda \in [0, 1] \) and
\[
\|y\| \leq \|\lambda x + (1 - \lambda)z\| \quad \text{for all } \lambda \in [0, 1].
\]
Show that the map \( A \) is closed for each of the following cases:

a. \( \|\cdot\| \) denotes the Euclidean norm; that is, \( \|g\| = (\sum_{i=1}^{n} g_i^2)^{1/2} \).

b. \( \|\cdot\| \) denotes the \( \ell_1 \) norm; that is, \( \|g\| = \sum_{i=1}^{n} |g_i| \).

c. \( \|\cdot\| \) denotes the sup norm; that is, \( \|g\| = \max_{1 \leq i \leq n} |g_i| \).

Let \( A: \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n \) be the point-to-set map defined as follows. Given \( x \in \mathbb{R}^n \) and \( z \in \mathbb{R} \), then \( y \in A(x, z) \) means that \( \|y - x\| \leq z \) and
\[
\|y\| \leq \|w\| \quad \text{for each } w \text{ satisfying } \|w - x\| \leq z.
\]
Show that the map \( A \) is closed for each of the norms specified in Exercise 7.7.

Let \( X \) and \( Y \) be nonempty closed sets in \( \mathbb{R}^p \) and \( \mathbb{R}^q \), respectively. Show that the point-to-set map \( A: X \to Y \) is closed if and only if the set \( Z = \{(x, y) : x \in X, y \in A(x)\} \) is closed.

Consider the map \( A \), where \( A(x) \) is the nonnegative square root of \( x \). Starting from any positive \( x \), show that the algorithm defined by the map \( A \) converges to \( \bar{x} = 1 \). [Hint: Let \( \alpha(x) = |x - 1| \).]

Let \( \lambda \) be a given scalar, and let \( f: \mathbb{R} \to \mathbb{R} \) be continuously differentiable. Let \( A: \mathbb{R} \to \mathbb{R} \) be the point-to-point map defined as follows:
\[
A(x) = \begin{cases} 
  x + \lambda & \text{if } f(x + \lambda) < f(x) \\
  x - \lambda & \text{if } f(x + \lambda) \geq f(x) \text{ and } f(x - \lambda) < f(x) \\
  x & \text{if } f(x + \lambda) \geq f(x) \text{ and } f(x - \lambda) \geq f(x).
\end{cases}
\]

a. Show that the map \( A \) is closed on
\[
\Lambda = \{x : f(x + \lambda) \neq f(x) \text{ and } f(x - \lambda) \neq f(x)\}.
\]
b. Starting from $x_1 = 2.5$ and letting $\lambda = 1$, apply the algorithm defined by the map $A$ to minimize $f(x) = 2x^2 - 3x$.

c. Let $\Omega = \{x : |x - \bar{x}| \leq \lambda\}$, where $df(\bar{x})/dx = 0$. Verify that if the sequence of points generated by the algorithm is contained in a compact set, it converges to a point in $\Omega$.

d. Is it possible that the point $\bar{x}$ in Part c is a local maximum or a saddle point?

[7.12] Let $A: X \rightarrow X$, where $X = \{x : x \geq 1/2\}$ and $A(x) = |\sqrt{x}|$. Use Zangwill's convergence theorem to verify that starting with any point in $X$, the (entire) sequence generated by $A$ converges. Define $\Omega$ and $\alpha$ explicitly for this case. What is the rate of convergence?

[7.13] The line search map $M: R^n \times R^n \rightarrow R^n$ defined below is frequently encountered in nonlinear programming algorithms. The vector $y \in M(x, d)$ if it solves the following problem, where $f: R^n \rightarrow R$:

Minimize $f(x + \lambda d)$
subject to $x + \lambda d \geq 0$
$\lambda \geq 0$.

To show that $M$ is not closed, a sequence $(x_k, d_k)$ converging to $(x, d)$ and a sequence $y_k \in M(x_k, d_k)$ converging to $y$ must be exhibited such that $y \notin M(x, d)$. Given that $x_1 = (1, 0)'$, let $x_{k+1}$ be the point on the circle $(x_1 - 1)^2 + (x_2 - 1)^2 = 1$ midway between $x_k$ and $(0,1)'$. Let the vector $d_k = (x_{k+1} - x_k)/\|x_{k+1} - x_k\|$.

Letting $f(x_1, x_2) = (x_1 + 2)^2 + (x_2 - 2)^2$, show that:

a. The sequence $\{x_k\}$ converges to $x = (0,1)'$.

b. The vectors $\{d_k\}$ converge to $d = (0,1)'$.

c. The sequence $\{y_k\}$ converges to $y = (0,1)'$.

d. The map $M$ is not closed at $(x, d)$.

[7.14] Let $f: R^n \rightarrow R$ be a differentiable function. Consider the following direction-finding map $D: R^n \rightarrow R^n \times R^n$ that gives the deflected negative gradient. Given $x \geq 0$, then $(x, d) \in D(x)$ means that

$$d_j = \begin{cases} \frac{-\partial f(x)}{\partial x_j} & \text{if } x_j > 0, \text{ or } x_j = 0 \text{ and } \frac{\partial f(x)}{\partial x_j} \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$
Show that $D$ is not closed. [Hint: Let $f(x_1, x_2) = x_1 - x_2$ and consider the sequence $\{x_k\}$ converging to $(0,1)^t$, where $x_k = (1/k, 1)^t$.]

[7.15] Let $f: \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Consider the composite map $A = MD$, where $D: \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ and $M: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ are defined as follows. Given $x \geq 0$, then $(x, d) \in D(x)$ means that

$$d_j = \begin{cases} \frac{-\partial f(x)}{\partial x_j} & \text{if } x_j > 0, \text{ or } x_j = 0 \text{ and } \frac{\partial f(x)}{\partial x_j} \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The vector $y \in M(x, d)$ means that $y = x + \lambda d$ for some $\lambda \geq 0$, where $\lambda$ solves the problem to minimize $f(x + \lambda d)$ subject to $x + \lambda d \geq 0$, $\lambda \geq 0$.

a. Find an optimal solution to the following problem using the KKT conditions:

Minimize $x_1^2 + x_2^2 - x_1 x_2 + 2x_1 + x_2$

subject to $x_1, x_2 \geq 0$.

b. Starting from the point $(2, 1)$, solve the problem in Part a using the algorithm defined by the algorithmic map $A$. Note that the algorithm converges to the optimal solution obtained in Part a.

c. Starting from the point $(0, 0.09, 0)$, solve the following problem credited to Wolfe [1972] using the algorithm defined by $A$:

Minimize $(4/3)(x_1^2 - x_1 x_2 + x_2^2)^{3/4} - x_3$

subject to $x_1, x_2, x_3 \geq 0$.

Note that the sequence generated converges to the point $(0, 0, \bar{x}_3)$, where $\bar{x}_3 = 0.3(1 + 0.5\sqrt{2})$. Using the KKT conditions, show that this point is not an optimal solution. Note that the algorithm converges to an optimal solution in Part b but not in Part c. This is because the map $A$ is not closed, as seen in Exercises 7.12 and 7.13.

[7.16] Let $f: \mathbb{R} \to \mathbb{R}$ be continuously differentiable. Consider the point-to-point map $A: \mathbb{R} \to \mathbb{R}$ defined as follows, where $f'(x) = df(x)/dx$:

$$A(x) = x - \frac{f(x)}{f'(x)} \quad \text{if } f'(x) \neq 0.$$ 

a. Show that $A$ is closed on the set $\Lambda = \{x : f'(x) \neq 0\}$.

b. Let $f(x) = x^2 - 2x - 3$, and apply the above algorithm starting from the point $x_1 = -5$. Note that the algorithm converges to $x = -1$, where $f(-1) = 0$. 

c. For the function defined by \( f(x) = x^2 - 1x^3 \), verify that starting from \( x_1 = 3/5 \), the algorithm does not converge to a point \( x \), where \( f(x) = 0 \).

d. The algorithm defined by the closed map \( A \) is sometimes used to find a point where \( f \) is equal to zero. In Part b the algorithm converged, whereas in Part c it did not. Discuss in reference to Theorem 7.2.3.

[7.17] In Theorem 7.3.5 we assumed that \( \det(D(x)) > \varepsilon > 0 \). Could this assumption be replaced by the following?

At each point \( x_k \) generated by the algorithm, the search directions \( d_1, \ldots, d_n \) generated by the algorithm are linearly independent.

[7.18] Let \( A: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \) be defined as follows. Given \( c, d \in \mathbb{R}^n \), \( k \in \mathbb{R} \), and a compact polyhedral set \( X \subseteq \mathbb{R}^n \), then \( \overline{\lambda} \in A(c, d) \) if \( \overline{\lambda} = \sup \{ \lambda : z(\lambda) \geq k \} \), where \( z(\lambda) = \min \{ (c + \lambda d)' x : x \in X \} \). Show that the point-to-point map \( A \) is closed at \((c, d)\).

[7.19] Let \( f: \mathbb{R}^n \to \mathbb{R} \) be a continuous function, and let \( I \) be a closed bounded interval in \( \mathbb{R} \). Let \( A: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \) be the point-to-set map defined as follows. Given \( x, d \in \mathbb{R}^n \), where \( d \neq 0 \), \( y \in A(x, d) \) means that \( y = x + \lambda d \) for some \( \lambda \in I \), and furthermore, \( f(y) \leq f(x + \lambda d) \) for each \( \lambda \in I \).

a. Show that \( A \) is closed at \((x, d)\).

b. Does the result hold true if \( d = 0 \)?

c. Does the result hold true if \( I \) is not bounded?

[7.20] Let \( X \) be a closed set in \( \mathbb{R}^n \), and let \( f: \mathbb{R}^n \to \mathbb{R} \) and \( \beta: \mathbb{R}^n \to \mathbb{R}^{m+\ell} \) be continuous. Show that the point-to-set map \( C: \mathbb{R}^{m+\ell} \to \mathbb{R}^n \) defined below is closed:

\[
y \in C(w) \text{ if } y \text{ solves the problem to minimize } f(x) + w^T \beta(x) \text{ subject to } x \in X.
\]

[7.21] This exercise introduces a unified approach to the class of cutting plane methods that are frequently used in nonlinear programming. We state the algorithm and then give the assumptions under which the algorithm converges. The symbol \( \mathcal{S} \) represents the collection of polyhedral sets in \( \mathbb{R}^P \), and \( \Omega \) is the nonempty solution set in \( \mathbb{R}^q \).
General Cutting Plane Algorithm

**Initialization Step**

Choose a nonempty polyhedral set \( Z \subseteq R^p \), let \( k = 1 \), and go to the Main Step.

**Main Step**

1. Given \( Z_k \), let \( w_k \in B(Z_k) \), where \( B: S \rightarrow R^q \). If \( w_k \in \Omega \), stop; otherwise, go to Step 2.
2. Let \( v_k \in C(w_k) \), where \( C: R^q \rightarrow R^r \). Let \( a: R^r \rightarrow R \) and \( b: R^r \rightarrow R^p \) be continuous functions, and let
   \[
   Z_{k+1} = Z_k \cap \{ x: a(v_k) + b'(v_k)x \geq 0 \}.
   \]
   Replace \( k \) by \( k + 1 \), and repeat Step 1.

**Convergence of the Cutting Plane Algorithm**

Under the following assumptions, either the algorithm stops in a finite number of steps at a point in \( \Omega \), or it generates an infinite sequence \( \{ w_k \} \) such that all of its accumulation points belong to \( \Omega \).

1. \( \{ w_k \} \) and \( \{ v_k \} \) are contained in compact sets in \( R^q \) and \( R^r \), respectively.
2. For each \( Z \), if \( w \in B(Z) \), \( w \in Z \).
3. \( C \) is a closed map.
4. Given \( w \not\in \Omega \) and \( Z \), where \( w \in B(Z) \), \( v \in C(w) \) implies that \( w \not\in \{ x: a(v) + b'(v)x \geq 0 \} \) and \( Z \cap \{ x: a(v) + b'(v)x \geq 0 \} = \emptyset \).

Prove the above convergence theorem. [Hint: Let \( \{ w_k \} \subseteq R^q \) and \( \{ v_k \} \subseteq R^r \) be convergent subsequences with limits \( w \) and \( v \). First, show that for any \( k \), we must have
   \[
   a(v_k) + b'(v_k)w_{\ell} \geq 0 \quad \text{for all} \quad \ell \geq k+1.
   \]
   Taking limits, show that \( a(v) + b'(v)w \geq 0 \). This inequality, together with Assumptions 3 and 4, imply that \( w \in \Omega \), because otherwise, a contradiction can be obtained.]

[7.22] Consider the dual cutting plane algorithm described in Section 6.4 for maximizing the dual function.

- a. Show that the dual cutting plane algorithm is a special form of the general cutting plane algorithm discussed in Exercise 7.21.
- b. Verify that Assumptions 1 through 4 of the convergence theorem stated in Exercise 7.21 hold true, so that the dual cutting plane algorithm converges to an optimal solution to the dual problem. (Hint: Referring to Exercise 7.20, note that the map \( C \) is closed.)
The Concept of an Algorithm

This exercise describes the cutting plane algorithm of Kelley [1960] for solving a problem of the following form, where \( g_i \) for \( i = 1, \ldots, m \) are convex:

\[
\begin{align*}
\text{Minimize} \quad & c^T x \\
\text{subject to} \quad & g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m \\
& Ax \leq b.
\end{align*}
\]

Kelley’s Cutting Plane Algorithm

Initialization Step

Let \( X_1 \) be a polyhedral set such that \( X_1 \supseteq \{x : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m\} \). Let \( Z_1 = X_1 \cap \{x : Ax \leq b\} \), let \( k = 1 \), and go to the Main Step.

Main Step

1. Solve the linear program to minimize \( c^T x \) subject to \( x \in Z_k \). Let \( x_k \) be an optimal solution. If \( g_i(x_k) \leq 0 \) for all \( i \), stop; \( x_k \) is an optimal solution. Otherwise, go to Step 2.
2. Let \( g_j(x_k) = \max_{1 \leq i \leq m} g_i(x_k) \), and let

\[
Z_{k+1} = Z_k \cap \{x : g_j(x_k) + \nabla g_j(x_k)^T(x - x_k) \leq 0\}.
\]

Replace \( k \) by \( k + 1 \), and repeat Step 1. [Obviously, \( \nabla g_i(x_k) \neq 0 \), because otherwise, \( g_j(x) \geq g_j(x_k) + \nabla g_j(x_k)^T(x - x_k) > 0 \) for all \( x \), implying that the problem is infeasible.]

a. Apply the above algorithm to solve the following problem:

\[
\begin{align*}
\text{Minimize} \quad & -3x_1 - 2x_2 \\
\text{subject to} \quad & -x_1^2 + x_2 + 1 \leq 0 \\
& 2x_1 + 3x_2 \leq 6 \\
& x_1, \quad x_2 \geq 0.
\end{align*}
\]

b. Show that Kelley’s algorithm is a special case of the general cutting plane algorithm of Exercise 7.21.

c. Show that the above algorithm converges to an optimal solution using the convergence theorem of Exercise 7.21.

d. Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) and \( Ax \leq b \). Show how the problem can be reformulated so that the above algorithm is applicable.

[Hint: Consider adding the constraint \( f(x) - z \leq 0 \).]
This exercise describes the supporting hyperplane method of Veinott [1967] for solving a problem of the following form, where $g_i$ for all $i$ are pseudoconvex and where $g_i(\hat{x}) < 0$ for $i = 1, ..., m$ for some point $\hat{x} \in \mathbb{R}^n$:

Minimize $c^t x$

subject to $g_i(x) \leq 0$ for $i = 1, ..., m$

$Ax \leq b.$

Veinott's Supporting Hyperplane Algorithm

Initialization Step

Let $X_1$ be a polyhedral set such that $X_1 \supseteq \{x : g_i(x) \leq 0 \text{ for } i = 1, ..., m\}$. Let $Z_1 = X_1 \cap \{x : Ax \leq b\}$, let $k = 1$, and go to the Main Step.

Main Step

1. Solve the linear program to minimize $c^t x$ subject to $x \in Z_k$. Let $x_k$ be an optimal solution. If $g_i(x_k) \leq 0$ for all $i$, stop; $x_k$ is an optimal solution to the original problem. Otherwise, go to Step 2.

2. Let $\bar{x}_k$ be the point on the line segment joining $x_k$ and $\hat{x}$, and on the boundary of the region $\{x : g_i(x) \leq 0 \text{ for } i = 1, ..., m\}$. Let $g_i(\bar{x}_k) = 0$, and let

$$Z_{k+1} = Z_k \cap \{x : \nabla g_i(\bar{x}_k)^t(x - \bar{x}_k) \leq 0\}.$$ 

Replace $k$ by $k + 1$, and repeat Step 1.

[Note that $\nabla g_i(\bar{x}_k) \neq 0$, because otherwise, by the pseudoconvexity of $g_j$ and since $g_j(\bar{x}_k) = 0$, it follows that $g_j(x) \geq 0$ for all $x$, contradicting the fact that $g_j(\hat{x}) < 0$.]

a. Apply the above algorithm to the problem given in Part a of Exercise 7.23.

b. Show that Veinott's algorithm is a special case of the general cutting plane algorithm of Exercise 7.21.

c. Show that the above algorithm converges to an optimal solution using the convergence theorem of Exercise 7.21.

(Note that the above algorithm can handle convex objective functions by reformulating the problem as in Part d of Exercise 7.23.)

Notes and References

The concept of closed maps is related to that of upper and lower semicontinuity. For a study of this subject, see Berge [1963], Hausdorff [1962], and Meyer
Hogan [1973d] studied properties of point-to-set maps from the standpoint of mathematical programming, where a number of different definitions and results are compared and integrated.

Using the notion of a closed map, Zangwill [1969] presents a unified treatment of the subject of convergence of nonlinear programming algorithms. Theorem 7.2.3, which is used to prove convergence of many algorithms, is credited to Zangwill. Polak [1970, 1971] presents several convergence theorems that are related to Theorem 7.2.3. Polak’s main theorem applies to a larger number of algorithms than that of Zangwill because of its weaker assumptions. Huard [1975] has proved convergence of some general nonlinear programming algorithms using the notion of closed maps. The results of Polak and Zangwill ensure that all accumulation points of the sequence of points generated by an algorithm belong to a solution set. However, convergence of the complete sequence is not generally guaranteed.

Under the stronger assumption of closedness of the algorithmic map everywhere, and using the concept of a fixed point, Meyer [1976] proved convergence of the complete sequence of iterates to a fixed point. The utility of the result is somewhat limited, however, because many algorithmic maps are not closed at solution points.

In order to apply Theorem 7.2.3 to prove convergence of a given algorithm, we must show closedness of the overall map. Theorem 7.3.2, where the algorithmic map is viewed as the composition of several maps, may be of use here. Another approach would be to prove convergence of the algorithm directly, even though the overall map may not be closed. Theorems 7.3.4 and 7.3.5 prove convergence for two classes of such algorithms. The first relates to an algorithm that can be viewed as the composition of two maps, one of which satisfies the assumptions of Theorem 7.2.3. The second relates to an algorithm that searches along linearly independent directions.

In Section 7.4, the subject of the speed or rate of convergence is introduced briefly. Ortega and Rheinboldt [1970] give a detailed treatment of \( q \)-and \( r \)-order convergence rates. The parameters \( p \) and \( \beta \) in Definition 7.4.1 determine the order and rate of stepwise convergence to an optimal solution as the solution point is approached. For a discussion of average convergence rates, see Luenberger [1973a/1984]. Of particular importance is the notion of superlinear convergence. A great deal of research has been directed to establish rates of convergence of nonlinear programming algorithms. See Luenberger [1973a/1984] and the Notes and References section at the end of Chapter 8.

There is a class of methods for solving nonlinear programming problems that use cutting planes. An example of such a procedure is given in Section 6.4. Zangwill [1969] presents a unified treatment of cutting plane algorithms. A general theorem showing convergence of such algorithms is presented in Exercise 7.21. Exercises 7.22, 7.23, and 7.24, respectively, deal with convergence of the dual cutting plane method, Kelley’s [1960] cutting plane algorithm, and Veinott’s [1967] supporting hyperplane algorithm.
Chapter 8 Unconstrained Optimization

Unconstrained optimization deals with the problem of minimizing or maximizing a function in the absence of any restrictions. In this chapter we discuss both the minimization of a function of one variable and a function of several variables. Even though most practical optimization problems have side restrictions that must be satisfied, the study of techniques for unconstrained optimization is important for several reasons. Many algorithms solve a constrained problem by converting it into a sequence of unconstrained problems via Lagrangian multipliers, as illustrated in Chapter 6, or via penalty and barrier functions, as discussed in Chapter 9. Furthermore, most methods proceed by finding a direction and then minimizing along this direction. This line search is equivalent to minimizing a function of one variable without constraints or with simple constraints, such as lower and upper bounds on the variables. Finally, several unconstrained optimization techniques can be extended in a natural way to provide and motivate solution procedures for constrained problems.

Following is an outline of the chapter.

Section 8.1: Line Search Without Using Derivatives We discuss several procedures for minimizing strictly quasiconvex functions of one variable without using derivatives. Uniform search, dichotomous search, the golden section method, and the Fibonacci method are covered.

Section 8.2: Line Search Using Derivatives Differentiability is assumed, and the bisection search method and Newton’s method are discussed.

Section 8.3: Some Practical Line Search Methods We describe the popular quadratic-fit line search method and present the Armijo rule for performing acceptable, inexact line searches.

Section 8.4: Closedness of the Line Search Algorithmic Map We show that the line search algorithmic map is closed, a property that is essential in convergence analyses. Readers who are not interested in convergence analyses may skip this section.

Section 8.5: Multidimensional Search Without Using Derivatives The cyclic coordinate method, the method of Hooke and Jeeves, and Rosenbrock’s method are discussed. Convergence of these methods is also established.

Section 8.6: Multidimensional Search Using Derivatives We develop the steepest descent method and the method of Newton and analyze their convergence properties.

Section 8.7: Modification of Newton’s Method: Levenberg–Marquardt and Trust Region Methods We describe different variants of Newton’s method based on the Levenberg–Marquardt and trust region methods,
which ensure the global convergence of Newton's method. We also discuss some insightful connections between these methods.

Section 8.8: Methods Using Conjugate Directions: Quasi-Newton and Conjugate Gradient Methods

The important concept of conjugacy is introduced. If the objective function is quadratic, then methods using conjugate directions are shown to converge in a finite number of steps. Various quasi-Newton/variable metric and conjugate gradient methods are covered based on the concept of conjugate directions, and their computational performance and convergence properties are discussed.

Section 8.9: Subgradient Optimization Methods

We introduce the extension of the steepest descent algorithm to that of minimizing convex, non-differentiable functions via subgradient-based directions. Variants of this technique that are related to conjugate gradient and variable metric methods are mentioned, and the crucial step of selecting appropriate step sizes in practice is discussed.

8.1 Line Search Without Using Derivatives

One-dimensional search is the backbone of many algorithms for solving a nonlinear programming problem. Many nonlinear programming algorithms proceed as follows. Given a point $x_k$, find a direction vector $d_k$ and then a suitable step size $\lambda_k$, yielding a new point $x_{k+1} = x_k + \lambda_k d_k$; the process is then repeated. Finding the step size $\lambda_k$ involves solving the subproblem to minimize $f(x_k + \lambda d_k)$, which is a one-dimensional search problem in the variable $\lambda$. The minimization may be over all real $\lambda$, nonnegative $\lambda$, or $\lambda$ such that $x_k + \lambda d_k$ is feasible.

Consider a function $\theta$ of one variable $\lambda$ to be minimized. One approach to minimizing $\theta$ is to set the derivative $\theta'$ equal to 0 and then solve for $\lambda$. Note, however, that $\theta$ is usually defined implicitly in terms of a function $f$ of several variables. In particular, given the vectors $x$ and $d$, $\theta(\lambda) = f(x + \lambda d)$. If $f$ is not differentiable, then $\theta$ will not be differentiable. If $f$ is differentiable, then $\theta'(\lambda) = d^t \nabla f(x + \lambda d)$. Therefore, to find a point $\lambda$ with $\theta'(\lambda) = 0$, we have to solve the equation $d^t \nabla f(x + \lambda d) = 0$, which is usually nonlinear in $\lambda$. Furthermore, $\lambda$ satisfying $\theta'(\lambda) = 0$ is not necessarily a minimum; it may be a local minimum, a local maximum, or even a saddle point. For these reasons, and except for some special cases, we avoid minimizing $\theta$ by letting its derivative be equal to zero. Instead, we resort to some numerical techniques for minimizing the function $\theta$.

In this section we discuss several methods that do not use derivatives for minimizing a function $\theta$ of one variable over a closed bounded interval. These methods fall under the categories of simultaneous line search and sequential line search problems. In the former case, the candidate points are determined \textit{a priori},
whereas in the sequential search, the values of the function at the previous iterations are used to determine the succeeding points.

**Interval of Uncertainty**

Consider the line search problem to minimize $\theta(\lambda)$ subject to $a \leq \lambda \leq b$. Since the exact location of the minimum of $\theta$ over $[a, b]$ is not known, this interval is called the *interval of uncertainty*. During the search procedure if we can exclude portions of this interval that do not contain the minimum, then the interval of uncertainty is reduced. In general, $[a, b]$ is called the *interval of uncertainty* if a minimum point $\lambda^*$ lies in $[a, b]$, although its exact value is not known.

Theorem 8.1.1 shows that if the function $\theta$ is strictly quasiconvex, then the interval of uncertainty can be reduced by evaluating $\theta$ at two points within the interval.

**8.1.1 Theorem**

Let $\theta : \mathbb{R} \to \mathbb{R}$ be strictly quasiconvex over the interval $[a, b]$. Let $\lambda, \mu \in [a, b]$ be such that $\lambda < \mu$. If $\theta(\lambda) > \theta(\mu)$, then $\theta(z) \geq \theta(\mu)$ for all $z \in [a, \lambda]$. If $\theta(\lambda) \leq \theta(\mu)$, then $\theta(z) \geq \theta(\lambda)$ for all $z \in (\mu, b]$.

**Proof**

Suppose that $\theta(\lambda) > \theta(\mu)$, and let $z \in [a, \lambda)$. By contradiction, suppose that $\theta(z) < \theta(\mu)$. Since $\lambda$ can be written as a convex combination of $z$ and $\mu$, and by the strict quasiconvexity of $\theta$, we have

$$\theta(\lambda) < \max\{\theta(z), \theta(\mu)\} = \theta(\mu),$$

contradicting $\theta(\lambda) > \theta(\mu)$. Hence, $\theta(z) \geq \theta(\mu)$. The second part of the theorem can be proved similarly.

From Theorem 8.1.1, under strict quasiconvexity if $\theta(\lambda) > \theta(\mu)$, the new interval of uncertainty is $[\lambda, b]$. On the other hand, if $\theta(\lambda) \leq \theta(\mu)$, the new interval of uncertainty is $[a, \mu]$. These two cases are illustrated in Figure 8.1.

Literature on nonlinear programming frequently uses the concept of *strict unimodality* of $\theta$ to reduce the interval of uncertainty (see Exercise 3.60). In this book we are using the equivalent concept of strict quasiconvexity. (See Exercises 3.57, 3.60, and 8.10 for definitions of various forms of unimodality and their relationships with different forms of quasiconvexity.)

We now present several procedures for minimizing a strictly quasiconvex function over a closed bounded interval by iteratively reducing the interval of uncertainty.
Example of a Simultaneous Search: Uniform Search

Uniform search is an example of simultaneous search, where we decide beforehand the points at which the functional evaluations are to be made. The interval of uncertainty $[a_l, b_l]$ is divided into smaller subintervals via the grid points $a_l + k\delta$ for $k = 1, \ldots, n$, where $b_l = a_l + (n+1)\delta$, as illustrated in Figure 8.2. The function $\theta$ is evaluated at each of the $n$ grid points. Let $\hat{\lambda}$ be a grid point having the smallest value of $\theta$. If $\theta$ is strictly quasiconvex, it follows that a minimum of $\theta$ lies in the interval $[\hat{\lambda} - \delta, \hat{\lambda} + \delta]$.

Choice of the Grid Length $\delta$

We see that the interval of uncertainty $[a_l, b_l]$ is reduced, after $n$ functional evaluations, to an interval of length $2\delta$. Noting that $n = [(b_l - a_l)/\delta] - 1$, if we desire a small final interval of uncertainty, then a large number $n$ of function evaluations must be made. One technique that is often used to reduce the computational effort is to utilize a large grid size first and then switch to a finer grid size.
Sequential Search

As may be expected, more efficient procedures that utilize the information generated at the previous iterations in placing the subsequent iterate can be devised. Here, we discuss the following sequential search procedures: dichotomous search, the golden section method, and the Fibonacci method.

Dichotomous Search

Consider \( \theta: \mathbb{R} \to \mathbb{R} \) to be minimized over the interval \([a_1, b_1]\). Suppose that \( \theta \) is strictly quasiconvex. Obviously, the smallest number of functional evaluations that is needed to reduce the interval of uncertainty is two. In Figure 8.3 we consider the location of the two points \( \lambda_1 \) and \( \mu_1 \). In Figure 8.3a, for \( \theta = \theta_1 \), note that \( \theta(\lambda_1) < \theta(\mu_1) \); and hence, by Theorem 8.1.1, the new interval of uncertainty is \([a_1, \mu_1]\). However, for \( \theta = \theta_2 \), note that \( \theta(\lambda_1) > \theta(\mu_1) \); hence, by Theorem 8.1.1 the new interval of uncertainty is \([\lambda_1, b_1]\). Thus, depending on the function \( \theta \), the length of the new interval of uncertainty is equal to \( \mu_1 - a_1 \) or \( b_1 - \lambda_1 \).

Note, however, that we do not know, \textit{a priori}, whether \( \theta(\lambda_1) < \theta(\mu_1) \) or \( \theta(\lambda_1) > \theta(\mu_1) \).* Thus, the \textit{optimal strategy} is to place \( \lambda_1 \) and \( \mu_1 \) in such a way as to guard against the worst possible outcome, that is, to minimize the maximum of \( \mu_1 - a_1 \) and \( b_1 - \lambda_1 \). This can be accomplished by placing \( \lambda_1 \) and \( \mu_1 \) at the midpoint of the interval \([a_1, b_1]\). If we do this, however, we would have only one trial point and would not be able to reduce the interval of uncertainty.

* If the equality \( \theta(\lambda_1) = \theta(\mu_1) \) is true, then the interval of uncertainty can be reduced further to \([\lambda_1, \mu_1]\). It may be noted, however, that exact equality is quite unlikely to occur in practice.
uncertainty. Therefore, as shown in Figure 8.3b, $\lambda_1$ and $\mu_1$ are placed symmetrically, each at a distance $\epsilon > 0$ from the midpoint. Here, $\epsilon > 0$ is a scalar that is sufficiently small so that the new length of uncertainty, $\epsilon + (b_1 - a_1)/2$, is close enough to the theoretical optimal value of $(b_1 - a_1)/2$ and, in the meantime, would make the functional evaluations $\theta(\lambda_1)$ and $\theta(\mu_1)$ distinguishable.

In dichotomous search, we place each of the first two observations, $\lambda_1$ and $\mu_1$, symmetrically at a distance $\epsilon$ from the midpoint $(a_1 + b_1)/2$. Depending on the values of $\theta$ at $\lambda_1$ and $\mu_1$, a new interval of uncertainty is obtained. The process is then repeated by placing two new observations.

Summary of the Dichotomous Search Method

Following is a summary of the dichotomous method for minimizing a strictly quasiconvex function $\theta$ over the interval $[a_1, b_1]$.

**Initialization Step** Choose the distinguishability constant, $2\epsilon > 0$, and the allowable final length of uncertainty, $\ell > 0$. Let $[a_1, b_1]$ be the initial interval of uncertainty, let $k = 1$, and go to the Main Step.

**Main Step**

1. If $b_k - a_k < \ell$, stop; the minimum point lies in the interval $[a_k, b_k]$. Otherwise, consider $\lambda_k$ and $\mu_k$ defined below, and go to Step 2.

\[
\lambda_k = \frac{a_k + b_k}{2} - \epsilon, \quad \mu_k = \frac{a_k + b_k}{2} + \epsilon.
\]

2. If $\theta(\lambda_k) < \theta(\mu_k)$, let $a_{k+1} = a_k$ and $b_{k+1} = \mu_k$. Otherwise, let $a_{k+1} = \lambda_k$ and $b_{k+1} = b_k$. Replace $k$ by $k + 1$, and go to Step 1.

Note that the length of uncertainty at the beginning of iteration $k + 1$ is given by

\[
(b_{k+1} - a_{k+1}) = \frac{1}{2^k} (b_1 - a_1) + 2\epsilon \left(1 - \frac{1}{2^k}\right).
\]

This formula can be used to determine the number of iterations needed to achieve the desired accuracy. Since each iteration requires two observations, the formula can also be used to determine the number of observations.

**Golden Section Method**

To compare the various line search procedures, the following reduction ratio will be of use:
length of interval of uncertainty after $v$ observations are taken
length of interval of uncertainty before taking the observations

Obviously, more efficient schemes correspond to small ratios. In dichotomous search, the reduction ratio above is approximately $(0.5)^{v/2}$. We now describe the more efficient golden section method for minimizing a strictly quasiconvex function, whose reduction ratio is given by $(0.618)^{v-1}$.

At a general iteration $k$ of the golden section method, let the interval of uncertainty be $[a_k, b_k]$. By Theorem 8.1.1, the new interval of uncertainty $[a_{k+1}, b_{k+1}]$ is given by $[\lambda_k, b_k]$ if $\theta(\lambda_k) > \theta(\mu_k)$ and by $[a_k, \mu_k]$ if $\theta(\lambda_k) \leq \theta(\mu_k)$. The points $\lambda_k$ and $\mu_k$ are selected such that the following hold true.

1. The length of the new interval of uncertainty $b_{k+1} - a_{k+1}$ does not depend on the outcome of the $k$th iteration, that is, on whether $\theta(\lambda_k) > \theta(\mu_k)$ or $\theta(\lambda_k) \leq \theta(\mu_k)$. Therefore, we must have $b_k - \lambda_k = \mu_k - a_k$. Thus, if $\lambda_k$ is of the form

$$\lambda_k = a_k + (1 - \alpha)(b_k - a_k),$$  

where $\alpha \in (0,1)$, $\mu_k$ must be of the form

$$\mu_k = a_k + \alpha(b_k - a_k)$$

so that

$$b_{k+1} - a_{k+1} = \alpha(b_k - a_k).$$

2. As $\lambda_{k+1}$ and $\mu_{k+1}$ are selected for the purpose of a new iteration, either $\lambda_{k+1}$ coincides with $\mu_k$ or $\mu_{k+1}$ coincides with $\lambda_k$. If this can be realized, then during iteration $k + 1$, only one extra observation is needed. To illustrate, consider Figure 8.4 and the following two cases.

Case 1: $\theta(\lambda_k) > \theta(\mu_k)$. In this case, $a_{k+1} = \lambda_k$ and $b_{k+1} = b_k$. To satisfy $\lambda_{k+1} = \mu_k$, and applying (8.1) with $k$ replaced by $k + 1$, we get

$$\mu_k = \lambda_{k+1} = a_{k+1} + (1 - \alpha)(b_{k+1} - a_{k+1}) = \lambda_k + (1 - \alpha)(b_k - \lambda_k).$$

Figure 8.4 Golden section rule.
Substituting the expressions of $\lambda_k$ and $\mu_k$ from (8.1) and (8.2) into the above equation, we get $\alpha^2 + \alpha - 1 = 0$.

Case 2: $\theta(\lambda_k) \leq \theta(\mu_k)$. In this case, $a_{k+1} = a_k$ and $b_{k+1} = \mu_k$. To satisfy $\mu_{k+1} = \lambda_k$, and applying (8.2) with $k$ replaced by $k+1$, we get

$$\lambda_k = \mu_{k+1} = a_{k+1} + \alpha(b_{k+1} - a_{k+1}) = a_k + \alpha(\mu_k - a_k).$$

Noting (8.1) and (8.2), the above equation gives $\alpha^2 + \alpha - 1 = 0$. The roots of the equation $\alpha^2 + \alpha - 1 = 0$ are $\alpha \approx 0.618$ and $\alpha \approx -1.618$. Since $\alpha$ must be in the interval (0, 1), then $\alpha \approx 0.618$. To summarize, if at iteration $k$, $\mu_k$ and $\lambda_k$ are chosen according to (8.1) and (8.2), where $\alpha = 0.618$, then the interval of uncertainty is reduced by a factor of 0.618. At the first iteration, two observations are needed at $\lambda_1$ and $\mu_1$, but at each subsequent iteration only one evaluation is needed, since either $\lambda_{k+1} = \mu_k$ or $\mu_{k+1} = \lambda_k$.

**Summary of the Golden Section Method**

Following is a summary of the golden section method for minimizing a strictly quasiconvex function over the interval $[a_1, b_1]$.

**Initialization Step**

Choose an allowable final length of uncertainty $\ell > 0$. Let $[a_1, b_1]$ be the initial interval of uncertainty, and let $\lambda_1 = a_1 + (1 - \alpha)(b_1 - a_1)$ and $\mu_1 = a_1 + \alpha(b_1 - a_1)$, where $\alpha = 0.618$. Evaluate $\theta(\lambda_1)$ and $\theta(\mu_1)$, let $k = 1$, and go to the Main Step.

**Main Step**

1. If $b_k - a_k < \ell$, stop; the optimal solution lies in the interval $[a_k, b_k]$. Otherwise, if $\theta(\lambda_k) > \theta(\mu_k)$, go to Step 2; and if $\theta(\lambda_k) \leq \theta(\mu_k)$, go to Step 3.

2. Let $a_{k+1} = \lambda_k$ and $b_{k+1} = b_k$. Furthermore, let $\lambda_{k+1} = \mu_k$, and let $\mu_{k+1} = a_{k+1} + \alpha(b_{k+1} - a_{k+1})$. Evaluate $\theta(\mu_{k+1})$ and go to Step 4.

3. Let $a_{k+1} = a_k$ and $b_{k+1} = \mu_k$. Furthermore, let $\mu_{k+1} = \lambda_k$, and let $\lambda_{k+1} = a_{k+1} + (1 - \alpha)(b_{k+1} - a_{k+1})$. Evaluate $\theta(\lambda_{k+1})$ and go to Step 4.

4. Replace $k$ by $k + 1$ and go to Step 1.

**8.1.2 Example**

Consider the following problem:

Minimize $\lambda^2 + 2\lambda$

subject to $-3 \leq \lambda \leq 5$. 
Clearly, the function \( \theta \) to be minimized is strictly quasiconvex, and the initial interval of uncertainty is of length 8. We reduce this interval of uncertainty to one whose length is at most 0.2. The first two observations are located at

\[
\lambda_1 = -3 + 0.382(8) = 0.056, \quad \mu_1 = -3 + 0.618(8) = 1.944.
\]

Note that \( \theta(\lambda_1) < \theta(\mu_1) \). Hence, the new interval of uncertainty is \([-3, 1.944]\]. The process is repeated, and the computations are summarized in Table 8.1. The values of \( \theta \) that are computed at each iteration are indicated by an asterisk. After eight iterations involving nine observations, the interval of uncertainty is \([-1.112, -0.936]\), so that the minimum can be estimated to be the midpoint \(-1.024\). Note that the true minimum is in fact \(-1.0\).

**Fibonacci Search**

The Fibonacci method is a line search procedure for minimizing a strictly quasiconvex function \( \theta \) over a closed bounded interval. Similar to the golden section method, the Fibonacci search procedure makes two functional evaluations at the first iteration and then only one evaluation at each of the subsequent iterations. However, the procedure differs from the golden section method in that the reduction of the interval of uncertainty varies from one iteration to another.

The procedure is based on the Fibonacci sequence \( \{F_n\} \), defined as follows:

\[
F_{v+1} = F_v + F_{v-1}, \quad v = 1, 2, \ldots \tag{8.3}
\]

\[
F_0 = F_1 = 1.
\]

The sequence is therefore 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, ... . At iteration \( k \), suppose that the interval of uncertainty is \([a_k, b_k]\). Consider the two points \( \lambda_k \) and \( \mu_k \) given below, where \( n \) is the total number of functional evaluations planned.

\[
\lambda_k = a_k + \frac{F_{n-k-1}}{F_{n-k+1}}(b_k - a_k), \quad k = 1, \ldots, n - 1 \tag{8.4}
\]
\[
\mu_k = a_k + \frac{F_{n-k}}{F_{n-k+1}} (b_k - a_k), \quad k = 1, \ldots, n-1. \tag{8.5}
\]

By Theorem 8.1.1, the new interval of uncertainty \([a_{k+1}, b_{k+1}]\) is given by \([\lambda_k, b_k]\) if \(\theta(\lambda_k) > \theta(\mu_k)\) and is given by \([a_k, \mu_k]\) if \(\theta(\lambda_k) \leq \theta(\mu_k)\). In the former case, noting (8.4) and letting \(v = n - k\) in (8.3), we get

\[
b_{k+1} - a_{k+1} = b_k - \lambda_k
\]

\[
= b_k - a_k - \frac{F_{n-k-1}}{F_{n-k+1}} (b_k - a_k)
\]

\[
= \frac{F_{n-k}}{F_{n-k+1}} (b_k - a_k). \tag{8.6}
\]

In the latter case, noting (8.5), we get

\[
b_{k+1} - a_{k+1} = \mu_k - a_k
\]

\[
= \frac{F_{n-k}}{F_{n-k+1}} (b_k - a_k). \tag{8.7}
\]

Thus, in either case, the interval of uncertainty is reduced by the factor \(F_{n-k}/F_{n-k+1}\).

We now show that at iteration \(k + 1\), either \(\lambda_{k+1} = \mu_k\) or \(\mu_{k+1} = \lambda_k\), so that only one functional evaluation is needed. Suppose that \(\theta(\lambda_k) > \theta(\mu_k)\). Then, by Theorem 8.1.1, \(a_{k+1} = \lambda_k\) and \(b_{k+1} = b_k\). Thus, applying (8.4) with \(k\) replaced by \(k + 1\), we get

\[
\lambda_{k+1} = a_{k+1} + \frac{F_{n-k-2}}{F_{n-k}} (b_{k+1} - a_{k+1})
\]

\[
= \lambda_k + \frac{F_{n-k-2}}{F_{n-k}} (b_k - \lambda_k).
\]

Substituting for \(\lambda_k\) from (8.4), we get

\[
\lambda_{k+1} = a_k + \frac{F_{n-k-1}}{F_{n-k+1}} (b_k - a_k) + \frac{F_{n-k-2}}{F_{n-k}} \left(1 - \frac{F_{n-k-1}}{F_{n-k+1}}\right) (b_k - a_k).
\]

Letting \(v = n - k\) in (8.3), it follows that \(1 - (F_{n-k-1}/F_{n-k+1}) = F_{n-k}/F_{n-k+1}\). Substituting in the above equation, we get

\[
\lambda_{k+1} = a_k + \frac{F_{n-k-1} + F_{n-k-2}}{F_{n-k+1}} (b_k - a_k).
\]

Now let \(v = n - k - 1\) in (8.3), and noting (8.5) it follows that
\[ \lambda_{k+1} = a_k + \frac{F_{n-k}}{F_{n-k+1}} (b_k - a_k) = \mu_k. \]

Similarly, if \( \theta(\lambda_k) \leq \theta(\mu_k) \), the reader can easily verify that \( \mu_{k+1} = \lambda_k \). Thus, in either case, only one observation is needed at iteration \( k+1 \).

To summarize, at the first iteration, two observations are made, and at each subsequent iteration, only one observation is necessary. Thus, at the end of iteration \( n-1 \), we have completed \( n-1 \) functional evaluations. Furthermore, for \( k = n - 1 \), it follows from (8.4) and (8.5) that \( \lambda_{n-1} = \mu_{n-1} = (1/2)(a_{n-1} + b_{n-1}) \).

Since either \( \lambda_{n-1} = \mu_{n-2} \) or \( \mu_{n-1} = \lambda_{n-2} \), theoretically no new observations are to be made at this stage. However, in order to reduce the interval of uncertainty further, the last observation is placed slightly to the right or to the left of the midpoint \( \lambda_{n-1} = \mu_{n-1} \), so that \( (1/2)(b_{n-1} - a_{n-1}) \) is the length of the final interval of uncertainty \([a_n, b_n]\).

Choosing the Number of Observations

Unlike the dichotomous search method and the golden section procedure, the Fibonacci method requires that the total number of observations \( n \) be chosen beforehand. This is because placement of the observations is given by (8.4) and (8.5) and, hence, is dependent on \( n \). From (8.6) and (8.7), the length of the interval of uncertainty is reduced at iteration \( k \) by the factor \( F_{n-k}/F_{n-k+1} \). Hence, at the end of \( n-1 \) iterations, where \( n \) total observations have been made, the length of the interval of uncertainty is reduced from \( b_1 - a_1 \) to \( b_n - a_n = (b_1 - a_1)/F_n \). Therefore, \( n \) must be chosen such that \( (b_1 - a_1)/F_n \) reflects the accuracy required.

Summary of the Fibonacci Search Method

The following is a summary of the Fibonacci search method for minimizing a strictly quasiconvex function over the interval \([a_1, b_1]\).

**Initialization Step** Choose an allowable final length of uncertainty \( \ell > 0 \) and a distinguishability constant \( \varepsilon > 0 \). Let \([a_1, b_1]\) be the initial interval of uncertainty, and choose the number of observations \( n \) to be taken such that \( F_n > (b_1 - a_1)/\ell \). Let \( \lambda_1 = a_1 + (F_{n-2}/F_n)(b_1 - a_1) \) and \( \mu_1 = a_1 + (F_{n-3}/F_n)(b_1 - a_1) \). Evaluate \( \theta(\lambda_1) \) and \( \theta(\mu_1) \), let \( k = 1 \), and go to the Main Step.

**Main Step**

1. If \( \theta(\lambda_k) > \theta(\mu_k) \), go to Step 2; and if \( \theta(\lambda_k) \leq \theta(\mu_k) \), go to Step 3.
2. Let \( a_{k+1} = \lambda_k \) and \( b_{k+1} = b_k \). Furthermore, let \( \lambda_{k+1} = \mu_k \), and let \( \mu_{k+1} = a_{k+1} + (F_{n-k-1}/F_{n-k})(b_{k+1} - a_{k+1}) \). If \( k = n - 2 \), go to Step 5; otherwise, evaluate \( \theta(\mu_{k+1}) \) and go to Step 4.

Unconstrained Optimization 353
3. Let $a_{k+1} = a_k$ and $b_{k+1} = \mu_k$. Furthermore, let $\mu_{k+1} = \lambda_k$, and let $\lambda_{k+1} = a_{k+1} + (F_{n-k-2} / F_{n-k})(b_{k+1} - a_{k+1})$. If $k = n - 2$, go to Step 5; otherwise, evaluate $\theta(\lambda_{k+1})$ and go to Step 4.

4. Replace $k$ by $k + 1$ and go to Step 1.

5. Let $\lambda_n = \lambda_{n-1}$ and $\mu_n = \lambda_{n-1} + \varepsilon$. If $\theta(\lambda_n) > \theta(\mu_n)$, let $a_n = \lambda_n$ and $b_n = \lambda_{n-1}$. Otherwise, if $\theta(\lambda_n) \leq \theta(\mu_n)$, let $a_n = a_{n-1}$ and $b_n = \lambda_n$.

Stop; the optimal solution lies in the interval $[a_n, b_n]$.

8.1.3 Example

Consider the following problem:

$$\text{Minimize } \lambda^2 + 2\lambda$$

subject to $-3 \leq \lambda \leq 5$.

Note that the function is strictly quasiconvex on the interval and that the true minimum occurs at $\lambda = -1$. We reduce the interval of uncertainty to one whose length is, at most, 0.2. Hence, we must have $F_n > 8/0.2 = 40$, so that $n = 9$. We adopt the distinguishability constant $\varepsilon = 0.01$.

The first two observations are located at

$$\lambda_1 = -3 + \frac{F_7}{F_9}(8) = 0.054545, \quad \mu_1 = -3 + \frac{F_8}{F_9}(8) = 1.945454.$$ 

Note that $\theta(\lambda_1) < \theta(\mu_1)$. Hence, the new interval of uncertainty is $[-3.000000, 1.945454]$. The process is repeated, and the computations are summarized in Table 8.2. The values of $\theta$ that are computed at each iteration are indicated by an asterisk. Note that at $k = 8$, $\lambda_k = \mu_k = \lambda_{k-1}$, so that no functional evaluations are needed at this stage. For $k = 9$, $\lambda_k = \mu_k = \lambda_{k-1} = -0.963636$ and $\mu_k = \lambda_k + \varepsilon = -0.953636$. Since $\theta(\mu_k) > \theta(\lambda_k)$, the final interval of uncertainty $[a_9, b_9]$ is $[-1.109091, -0.963636]$, whose length $\ell$ is 0.145455. We approximate the minimum to be the midpoint $-1.036364$. Note from Example 8.1.2 that with the same number of observations $n = 9$, the golden section method gave a final interval of uncertainty whose length is 0.176.

Comparison of Derivative-Free Line Search Methods

Given a function $\theta$ that is strictly quasiconvex on the interval $[a_1, b_1]$, obviously each of the methods discussed in this section will yield a point $\lambda$ in a finite number of steps such that $|\lambda - \lambda| \leq \ell$, where $\ell$ is the length of the final interval of uncertainty and $\lambda$ is the minimum point over the interval. In particular, given the length $\ell$ of the final interval of uncertainty, which reflects the desired degree
Table 8.2 Summary of Computations for the Fibonacci Search Method

<table>
<thead>
<tr>
<th>Iteration</th>
<th>( a_k )</th>
<th>( b_k )</th>
<th>( \lambda_k )</th>
<th>( \mu_k )</th>
<th>( \theta(\lambda_k) )</th>
<th>( \theta(\mu_k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-3.000000</td>
<td>5.000000</td>
<td>0.054545</td>
<td>1.945454</td>
<td>0.112065*</td>
<td>7.675699*</td>
</tr>
<tr>
<td>2</td>
<td>-3.000000</td>
<td>1.945454</td>
<td>-1.109091</td>
<td>0.054545</td>
<td>-0.988099*</td>
<td>0.112065</td>
</tr>
<tr>
<td>3</td>
<td>-3.000000</td>
<td>0.054545</td>
<td>-1.836363</td>
<td>-1.109091</td>
<td>-0.30497*</td>
<td>-0.988099</td>
</tr>
<tr>
<td>4</td>
<td>-1.836363</td>
<td>0.054545</td>
<td>-1.109091</td>
<td>-0.672727</td>
<td>-0.988099</td>
<td>-0.892892*</td>
</tr>
<tr>
<td>5</td>
<td>-1.836363</td>
<td>-0.672727</td>
<td>-1.399999</td>
<td>-1.109091</td>
<td>-0.840001*</td>
<td>-0.988099</td>
</tr>
<tr>
<td>6</td>
<td>-1.399999</td>
<td>-0.672727</td>
<td>-1.109091</td>
<td>-0.963636</td>
<td>-0.988099</td>
<td>-0.998677*</td>
</tr>
<tr>
<td>7</td>
<td>-1.109091</td>
<td>-0.672727</td>
<td>-0.963636</td>
<td>-0.818182</td>
<td>-0.998677</td>
<td>-0.966942*</td>
</tr>
<tr>
<td>8</td>
<td>-1.109091</td>
<td>-0.818182</td>
<td>-0.963636</td>
<td>-0.963636</td>
<td>-0.998677</td>
<td>-0.998677</td>
</tr>
<tr>
<td>9</td>
<td>-1.109091</td>
<td>-0.963636</td>
<td>-0.963636</td>
<td>-0.953636</td>
<td>-0.998677</td>
<td>-0.997850*</td>
</tr>
</tbody>
</table>

of accuracy, the required number of observations \( n \) can be computed as the smallest positive integer satisfying the following relationships.

Uniform search method: \[ n \geq \frac{b_1 - a_1}{\ell/2} - 1. \]

Dichotomous search method: \[ (1/2)^{n/2} \leq \frac{\ell}{b_1 - a_1}. \]

Golden section method: \[ (0.618)^{n-1} \leq \frac{\ell}{b_1 - a_1}. \]

Fibonacci search method: \[ F_n \geq \frac{b_1 - a_1}{\ell}. \]

From the above expressions, we see that the number of observations needed is a function of the ratio \((b_1 - a_1)/\ell\). Hence, for a fixed ratio \((b_1 - a_1)/\ell\), the smaller the number of observations required, the more efficient is the algorithm. It should be evident that the most efficient algorithm is the Fibonacci method, followed by the golden section procedure, the dichotomous search method, and finally the uniform search method.

Also note that for \( n \) large enough, \( 1/F_n \) is asymptotic to \((0.618)^{n-1}\), so that the Fibonacci search method and the golden section are almost identical. It is worth mentioning that among the derivative-free methods that minimize strict quasiconvex functions over a closed bounded interval, the Fibonacci search method is the most efficient in that it requires the smallest number of observations for a given reduction in the length of the interval of uncertainty.

**General Functions**

The procedures discussed above all rely on the strict quasiconvexity assumption. In many problems this assumption does not hold true, and in any case, it cannot
be verified easily. One way to handle this difficulty, especially if the initial interval of uncertainty is large, is to divide it into smaller intervals, find the minimum over each subinterval and choose the smallest of the minima over the subintervals. (A more refined global optimization scheme could also be adopted; see the Notes and References section.) Alternatively, one can simply apply the method assuming strict quasiconvexity and allow the procedure to converge to some local minimum solution.

8.2 Line Search Using Derivatives

In the preceding section we discussed several line search procedures that use functional evaluations. In this section we discuss the bisection search method and Newton's method, both of which need derivative information.

Bisection Search Method

Suppose that we wish to minimize a function \( \theta \) over a closed and bounded interval. Furthermore, suppose that \( \theta \) is pseudoconvex and hence, differentiable. At iteration \( k \), let the interval of uncertainty be \([a_k, b_k]\). Suppose that the derivative \( \theta'(\lambda_k) \) is known, and consider the following three possible cases:

1. If \( \theta'(\lambda_k) = 0 \), then, by the pseudoconvexity of \( \theta \), \( \lambda_k \) is a minimizing point.

2. If \( \theta'(\lambda_k) > 0 \), then, for \( \lambda > \lambda_k \) we have \( \theta'(\lambda_k)(\lambda - \lambda_k) > 0 \); and by the pseudoconvexity of \( \theta \) it follows that \( \theta(\lambda) \geq \theta(\lambda_k) \). In other words, the minimum occurs to the left of \( \lambda_k \), so that the new interval of uncertainty \([a_{k+1}, b_{k+1}]\) is given by \([a_k, \lambda_k]\).

3. If \( \theta'(\lambda_k) < 0 \), then, for \( \lambda < \lambda_k \), \( \theta'(\lambda_k)(\lambda - \lambda_k) > 0 \), so that \( \theta(\lambda) \geq \theta(\lambda_k) \). Thus, the minimum occurs to the right of \( \lambda_k \), so that the new interval of uncertainty \([a_{k+1}, b_{k+1}]\) is given by \([\lambda_k, b_k]\).

The position of \( \lambda_k \) in the interval \([a_k, b_k]\) must be chosen so that the maximum possible length of the new interval of uncertainty is minimized. That is, \( \lambda_k \) must be chosen so as to minimize the maximum of \( \lambda_k - a_k \) and \( b_k - \lambda_k \). Obviously, the optimal location of \( \lambda_k \) is the midpoint \((1/2)(a_k + b_k)\).

To summarize, at any iteration \( k \), \( \theta' \) is evaluated at the midpoint of the interval of uncertainty. Based on the value of \( \theta' \), we either stop or construct a new interval of uncertainty whose length is half that at the previous iteration. Note that this procedure is very similar to the dichotomous search method except that at each iteration, only one derivative evaluation is required, as opposed to two functional evaluations for the dichotomous search method. However, the latter is akin to a finite difference derivative evaluation.
Convergence of the Bisection Search Method

Note that the length of the interval of uncertainty after \( n \) observations is equal to 
\[
(1/2)^n (b_1 - a_1),
\]
so that the method converges to a minimum point within any desired degree of accuracy. In particular, if the length of the final interval of uncertainty is fixed at \( \ell \), then \( n \) must be chosen to be the smallest integer such that 
\[
(1/2)^n \leq \ell / (b_1 - a_1).
\]

Summary of the Bisection Search Method

We now summarize the bisection search procedure for minimizing a pseudoconvex function \( \theta \) over a closed and bounded interval.

**Initialization Step** Let \([a_1, b_1] \) be the initial interval of uncertainty, and let \( \ell \) be the allowable final interval of uncertainty. Let \( n \) be the smallest positive integer such that 
\[
(1/2)^n \leq \ell / (b_1 - a_1). \]
Let \( k = 1 \) and go to the Main Step.

**Main Step**

1. Let \( \lambda_k = (1/2)(a_k + b_k) \) and evaluate \( \theta'(\lambda_k) \). If \( \theta'(\lambda_k) = 0 \), stop; \( \lambda_k \)
is an optimal solution. Otherwise, go to Step 2 if \( \theta'(\lambda_k) > 0 \), and go to Step 3 if \( \theta'(\lambda_k) < 0 \).
2. Let \( a_{k+1} = a_k \) and \( b_{k+1} = \lambda_k \). Go to Step 4.
3. Let \( a_{k+1} = \lambda_k \) and \( b_{k+1} = b_k \). Go to Step 4.
4. If \( k = n \), stop; the minimum lies in the interval \([a_{n+1}, b_{n+1}]\). Otherwise, replace \( k \) by \( k + 1 \) and repeat Step 1.

8.2.1 Example

Consider the following problem:

Minimize \( \lambda^2 + 2\lambda \)
subject to \(-3 \leq \lambda \leq 6.\)

Suppose that we want to reduce the interval of uncertainty to an interval whose length \( \ell \) is less than or equal to 0.2. Hence, the number of observations \( n \) satisfying 
\[
(1/2)^n \leq \ell / (b_1 - a_1) = 0.2/9 = 0.0222
\]
is given by \( n = 6 \). A summary of the computations using the bisection search method is given in Table 8.3. Note that the final interval of uncertainty is \([-1.0313, -0.8907] \), so that the minimum could be taken as the midpoint, \(-0.961\).
### Table 8.3 Summary of Computations for the Bisection Search Method

<table>
<thead>
<tr>
<th>Iteration $k$</th>
<th>$a_k$</th>
<th>$b_k$</th>
<th>$\lambda_k$</th>
<th>$\theta'(\lambda_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-3.0000</td>
<td>6.0000</td>
<td>1.5000</td>
<td>5.0000</td>
</tr>
<tr>
<td>2</td>
<td>-3.0000</td>
<td>1.5000</td>
<td>-0.7500</td>
<td>0.5000</td>
</tr>
<tr>
<td>3</td>
<td>-3.0000</td>
<td>-0.7500</td>
<td>-1.8750</td>
<td>-1.7500</td>
</tr>
<tr>
<td>4</td>
<td>-1.8750</td>
<td>-0.7500</td>
<td>-1.3125</td>
<td>-0.6250</td>
</tr>
<tr>
<td>5</td>
<td>-1.3125</td>
<td>-0.7500</td>
<td>-1.0313</td>
<td>-0.0625</td>
</tr>
<tr>
<td>6</td>
<td>-1.0313</td>
<td>-0.7500</td>
<td>-0.8907</td>
<td>0.2186</td>
</tr>
<tr>
<td>7</td>
<td>-1.0313</td>
<td>-0.8907</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Newton's Method

Newton's method is based on exploiting the quadratic approximation of the function $\theta$ at a given point $\lambda_k$. This quadratic approximation $q$ is given by

$$q(\lambda) = \theta(\lambda_k) + \theta'(\lambda_k)(\lambda - \lambda_k) + \frac{1}{2} \theta''(\lambda_k)(\lambda - \lambda_k)^2.$$  

The point $\lambda_{k+1}$ is taken to be the point where the derivative of $q$ is equal to zero. This yields $\theta'(\lambda_k) + \theta''(\lambda_k)(\lambda_{k+1} - \lambda_k) = 0$, so that

$$\lambda_{k+1} = \lambda_k - \frac{\theta'(\lambda_k)}{\theta''(\lambda_k)}.$$  

(8.8)

The procedure is terminated when $|\lambda_{k+1} - \lambda_k| < \varepsilon$, or when $|\theta'(\lambda_k)| < \varepsilon$, where $\varepsilon$ is a prespecified termination scalar.

Note that the above procedure can only be applied for twice differentiable functions. Furthermore, the procedure is well defined only if $\theta''(\lambda_k) \neq 0$ for each $k$.

#### 8.2.2 Example

Consider the function $\theta$:

$$\theta(\lambda) = \begin{cases} 
4\lambda^3 - 3\lambda^4 & \text{if } \lambda \geq 0 \\
4\lambda^3 + 3\lambda^4 & \text{if } \lambda < 0.
\end{cases}$$

Note that $\theta$ is twice differentiable everywhere. We apply Newton's method, starting from two different points. In the first case, $\lambda_1 = 0.40$; and as shown in Table 8.4, the procedure produces the point $0.002807$ after six iterations. The reader can verify that the procedure indeed converges to the stationary point $\lambda = 0$. In the second case, $\lambda_1 = 0.60$, and the procedure oscillates between the points 0.60 and -0.60, as shown in Table 8.5.
Unconstrained Optimization

Table 8.4 Summary of Computations for Newton's Method Starting from \( \lambda_1 = 0.4 \)

<table>
<thead>
<tr>
<th>Iteration ( k )</th>
<th>( \lambda_k )</th>
<th>( \theta'(\lambda_k) )</th>
<th>( \theta''(\lambda_k) )</th>
<th>( \lambda_{k+1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.400000</td>
<td>1.152000</td>
<td>3.840000</td>
<td>0.100000</td>
</tr>
<tr>
<td>2</td>
<td>0.100000</td>
<td>0.108000</td>
<td>2.040000</td>
<td>0.047059</td>
</tr>
<tr>
<td>3</td>
<td>0.047059</td>
<td>0.025324</td>
<td>1.049692</td>
<td>0.022934</td>
</tr>
<tr>
<td>4</td>
<td>0.022934</td>
<td>0.006167</td>
<td>0.531481</td>
<td>0.011331</td>
</tr>
<tr>
<td>5</td>
<td>0.11331</td>
<td>0.001523</td>
<td>0.267322</td>
<td>0.005634</td>
</tr>
<tr>
<td>6</td>
<td>0.005634</td>
<td>0.000379</td>
<td>0.134073</td>
<td>0.002807</td>
</tr>
</tbody>
</table>

Convergence of Newton's Method

The method of Newton, in general, does not converge to a stationary point starting with an arbitrary initial point. Observe that, in general, Theorem 7.2.3 cannot be applied as a result of the unavailability of a descent function. However, as shown in Theorem 8.2.3, if the starting point is sufficiently close to a stationary point, then a suitable descent function can be devised so that the method converges.

8.2.3 Theorem

Let \( \theta : R \to R \) be continuously twice differentiable. Consider Newton's algorithm defined by the map \( A(\lambda) = \lambda - \theta'(\lambda) / \theta''(\lambda) \). Let \( \lambda^* \) be such that \( \theta'(\lambda^*) = 0 \) and \( \theta''(\lambda^*) \neq 0 \). Let the starting point \( \lambda_1 \) be sufficiently close to \( \lambda^* \) so that there exist scalars \( k_1, k_2 > 0 \) with \( k_1 k_2 < 1 \) such that

1. \( \frac{1}{|\theta''(\lambda)|} \leq k_1 \)
2. \( \frac{|\theta(\lambda) - \theta'(\lambda) - \theta''(\lambda)(\lambda - \lambda^*)|}{(\lambda - \lambda^*)} \leq k_2 \)

for each \( \lambda \) satisfying \( |\lambda - \lambda^*| \leq |\lambda_1 - \lambda^*| \). Then the algorithm converges to \( \lambda^* \).

Table 8.5 Summary of Computations for Newton's Method Starting from \( \lambda_1 = 0.6 \)

<table>
<thead>
<tr>
<th>Iteration ( k )</th>
<th>( \lambda_k )</th>
<th>( \theta'(\lambda_k) )</th>
<th>( \theta''(\lambda_k) )</th>
<th>( \lambda_{k+1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.600</td>
<td>1.728</td>
<td>1.440</td>
<td>-0.600</td>
</tr>
<tr>
<td>2</td>
<td>-0.600</td>
<td>1.728</td>
<td>-1.440</td>
<td>0.600</td>
</tr>
<tr>
<td>3</td>
<td>0.600</td>
<td>1.728</td>
<td>1.440</td>
<td>-0.600</td>
</tr>
<tr>
<td>4</td>
<td>-0.600</td>
<td>1.728</td>
<td>-1.440</td>
<td>0.600</td>
</tr>
</tbody>
</table>
Proof

Let the solution set \( \Omega = \{ \bar{\lambda} \} \), and let \( X = \{ \lambda : |\lambda - \bar{\lambda}| \leq |\lambda_1 - \bar{\lambda}| \} \). We prove convergence by using Theorem 7.2.3. Note that \( X \) is compact and that the map \( A \) is closed on \( X \). We now show that \( \alpha(\lambda) = |\lambda - \bar{\lambda}| \) is indeed a descent function. Let \( \lambda \in X \) and suppose that \( \lambda \neq \bar{\lambda} \). Let \( \hat{\lambda} \in A(\lambda) \). Then, by the definition of \( A \) and since \( \theta'(\bar{\lambda}) = 0 \), we get

\[
\hat{\lambda} - \bar{\lambda} = (\lambda - \bar{\lambda}) - \frac{1}{\theta^*(\lambda)}[\theta'(\lambda) - \theta'(\bar{\lambda})]
\]

\[
= \frac{1}{\theta^*(\lambda)}[\theta'(\bar{\lambda}) - \theta'(\lambda) - \theta^*(\lambda)(\bar{\lambda} - \lambda)].
\]

Noting the hypothesis of the theorem, it then follows that

\[
|\hat{\lambda} - \bar{\lambda}| = \frac{1}{|\theta^*(\lambda)|} \frac{\theta'(\bar{\lambda}) - \theta'(\lambda) - \theta^*(\lambda)(\bar{\lambda} - \lambda)}{|\bar{\lambda} - \lambda|} |\lambda - \bar{\lambda}| \leq k_1 k_2 |\lambda - \bar{\lambda}| < |\lambda - \bar{\lambda}|.
\]

Therefore, \( \alpha \) is indeed a descent function, and the result follows immediately by the corollary to Theorem 7.2.3.

8.3 Some Practical Line Search Methods

In the preceding two sections we presented various line search methods that either use or do not use derivative-based information. Of these, the golden section method (which is a limiting form of Fibonacci’s search method) and the bisection method are often applied in practice, sometimes in combination with other methods. However, these methods follow a restrictive pattern of placing subsequent observations and do not accelerate the process by adaptively exploiting information regarding the shape of the function. Although Newton’s method tends to do this, it requires second-order derivative information and is not globally convergent. The quadratic-fit technique described in the discussion that follows adopts this philosophy, enjoys global convergence under appropriate assumptions such as pseudoconvexity, and is a very popular method.

We remark here that quite often in practice, whenever ill-conditioning effects are experienced with this method or if it fails to make sufficient progress during an iteration, a switchover to the bisection search procedure is typically made. Such a check for a possible switchover is referred to as a safeguard technique.
Quadratic-Fit Line Search

Suppose that we are trying to minimize a continuous, strictly quasiconvex function $\theta(\lambda)$ over $\lambda \geq 0$, and assume that we have three points $0 \leq \lambda_1 < \lambda_2 < \lambda_3$ such that $\theta_1 \geq \theta_2$ and $\theta_2 \leq \theta_3$, where $\theta_j = \theta(\lambda_j)$ for $j = 1, 2, 3$. Note that if $\theta_1 = \theta_2 = \theta_3$, then, by the nature of $\theta$, it is easily verified that these must all be minimizing solutions (see Exercise 8.12). Hence, suppose that in addition, at least one of the inequalities $\theta_1 > \theta_2$ and $\theta_2 < \theta_3$ holds true. Let us refer to the conditions satisfied by these three points as the three-point pattern (TPP). To begin with, we can take $\lambda_1 = 0$ and examine a trial point $\hat{\lambda}$, which might be the step length of a line search at the previous iteration of an algorithm. Let $\hat{\theta} = \theta(\hat{\lambda})$. If $\hat{\theta} \geq \theta_1$, we can set $\lambda_3 = \hat{\lambda}$ and find $\lambda_2$ by repeatedly halving the interval $[\lambda_1, \lambda_3]$ until a TPP is obtained. On the other hand, if $\hat{\theta} < \theta_1$, we can set $\lambda_2 = \hat{\lambda}$ and find $\lambda_3$ by doubling the interval $[\lambda_1, \lambda_2]$ until a TPP is obtained.

Now, given the three points $(\lambda_j, \theta_j)$, $j = 1, 2, 3$, we can fit a quadratic curve passing through these points and find its minimizer $\bar{\lambda}$, which must lie in $(\lambda_1, \lambda_3)$ by the TPP (see Exercise 8.11). There are three cases to consider. Denote $\bar{\theta} = \theta(\bar{\lambda})$ and let $\lambda_{\text{new}}$ denote the revised set of three points $(\lambda_1, \lambda_2, \lambda_3)$ found as follows:

**Case 1:** $\bar{\lambda} > \lambda_2$ (see Figure 8.5). If $\bar{\theta} \geq \theta_2$, then we let $\lambda_{\text{new}} = (\lambda_1, \lambda_2, \bar{\lambda})$. On the other hand, if $\bar{\theta} \leq \theta_2$, we let $\lambda_{\text{new}} = (\lambda_2, \bar{\lambda}, \lambda_3)$. (Note that in case $\bar{\theta} = \theta_2$, either choice is permissible.)

**Case 2:** $\bar{\lambda} < \lambda_2$. Similar to Case 1, if $\bar{\theta} \geq \theta_2$, we let $\lambda_{\text{new}} = (\bar{\lambda}, \lambda_2, \lambda_3)$; and if $\bar{\theta} \leq \theta_2$, we let $\lambda_{\text{new}} = (\lambda_1, \bar{\lambda}, \lambda_2)$.

![Figure 8.5 Quadratic-fit line search.](image-url)
Case 3: $\bar{\lambda} = \lambda_2$. In this case, we do not have a distinct point to obtain a new TPP. If $\lambda_3 - \lambda_1 \leq \varepsilon$ for some convergence tolerance $\varepsilon > 0$, we stop with $\lambda_2$ as the prescribed step length. Otherwise, we place a new observation point $\bar{\lambda}$ at a distance $\varepsilon/2$ away from $\lambda_2$ toward $\lambda_1$ or $\lambda_3$, whichever is further. This yields the situation described by Case 1 or 2 above, and hence, a new set of points defining $\lambda_{\text{new}}$ may be obtained accordingly.

Again, with respect to $\lambda_{\text{new}}$ if $\theta_1 = \theta_2 = \theta_3$ or if $(\lambda_3 - \lambda_1) \leq \varepsilon$ [or if $\theta'(\lambda_2) = 0$ in the differentiable case, or if some other termination criterion such as an acceptable step length in an inexact line search as described next below holds true], then we terminate this process. Otherwise, $\lambda_{\text{new}}$ satisfies the TPP, and the above procedure can be repeated using this new TPP.

Note that in Case 3 of the above procedure, when $\bar{\lambda} = \lambda_2$ the step of placing an observation in the vicinity of $\lambda_2$ is akin to evaluating $\theta'(\lambda_2)$ when $\theta$ is differentiable. In fact, if we assume that $\theta$ is pseudoconvex and continuously twice differentiable, and we apply a modified version of the foregoing procedure that uses derivatives to represent limiting cases of coincident observation values as described in Exercise 8.13, we can use Theorem 7.2.3 to demonstrate convergence to an optimal solution, given a starting solution $(\lambda_1, \lambda_2, \lambda_3)$ that satisfies the TPP.

Inexact Line Searches: Armijo’s Rule

Very often in practice, we cannot afford the luxury of performing an exact line search because of the expense of excessive function evaluations, even if we terminate with some small accuracy tolerance $\varepsilon > 0$. On the other hand, if we sacrifice accuracy, we might impair the convergence of the overall algorithm that iteratively employs such a line search. However, if we adopt a line search that guarantees a sufficient degree of accuracy or descent in the function value in a well-defined sense, this might induce the overall algorithm to converge. Below we describe one popular definition of an acceptable step length known as Armijo’s rule and refer the reader to the Notes and References section and Exercise 8.8 for other such exact line search criteria.

Armijo’s rule is driven by two parameters, $0 < \varepsilon < 1$ and $\alpha > 1$, which manage the acceptable step length from being too large or too small, respectively. (Typical values are $\varepsilon = 0.2$ and $\alpha = 2$.) Suppose that we are minimizing some differentiable function $f: \mathbb{R}^n \to \mathbb{R}$ at the point $\bar{x} \in \mathbb{R}^n$ in the direction $d \in \mathbb{R}^n$, where $\nabla f(\bar{x})^t d < 0$. Hence, $d$ is a descent direction. Define the line search function $\theta: \mathbb{R} \to \mathbb{R}$ as $\theta(\lambda) = f(\bar{x} + \lambda d)$ for $\lambda \geq 0$. Then the first-order approximation of $\theta$ at $\lambda = 0$ is given by $\theta(0) + \lambda \theta'(0)$ and is depicted in Figure 8.6. Now define
Unconstrained Optimization

Figure 8.6 Armijo’s rule.

\[ \hat{\theta}(\lambda) = \theta(0) + \lambda \theta'(0) \quad \text{for } \lambda \geq 0. \]

A step length \( \lambda \) is considered to be acceptable provided that \( \theta(\lambda) \leq \hat{\theta}(\lambda) \).

However, to prevent \( \lambda \) from being too small, Armijo’s rule also requires that \( \theta(\alpha \lambda) > \hat{\theta}(\alpha \lambda) \).

This gives an acceptable range for \( \lambda \), as shown in Figure 8.6.

Frequently, Armijo’s rule is adopted in the following manner. A fixed-step-length parameter \( \lambda \) is chosen. If \( \theta(\lambda) \leq \hat{\theta}(\lambda) \), then either \( \lambda \) is itself selected as the step size, or \( \lambda \) is sequentially doubled (assuming that \( \alpha = 2 \)) to find the largest integer \( t \geq 0 \) for which \( \theta(2^t \lambda) \leq \hat{\theta}(2^t \lambda) \). On the other hand, if \( \theta(\lambda) > \hat{\theta}(\lambda) \), then \( \lambda \) is sequentially halved to find the smallest integer \( t \geq 1 \) for which \( \theta(\lambda/2^t) \leq \hat{\theta}(\lambda/2^t) \).

Later, in Section 8.6, we analyze the convergence of a steepest descent algorithm that employs such a line search criterion.

8.4 Closedness of the Line Search Algorithmic Map

In the preceding three sections we discussed several procedures for minimizing a function of one variable. Since the one-dimensional search is a component of most nonlinear programming algorithms, we show in this section that line search procedures define a closed map.

Consider the line search problem to minimize \( \theta(\lambda) \) subject to \( \lambda \in L \), where \( \theta(\lambda) = f(x + \lambda d) \) and \( L \) is a closed interval in \( R \). This line search problem can be defined by the algorithmic map \( M: R^n \times R^n \to R^n \), defined by
\[ M(x, d) = \{ y : y = x + \lambda d \text{ for some } \lambda \in L \text{ and } f(y) \leq f(x + \lambda d) \text{ for each } \lambda \in L \}. \]

Note that \( M \) is generally a point-to-set map because there can be more than one minimizing point \( y \). Theorem 8.4.1 shows that the map \( M \) is closed. Thus, if the map \( D \) that determines the direction \( d \) is also closed, then, by Theorem 7.3.2 or its corollaries, if the additional conditions stated hold true, the overall algorithmic map \( A = MD \) is closed.

8.4.1 Theorem

Let \( f : \mathbb{R}^n \to \mathbb{R} \), and let \( L \) be a closed interval in \( \mathbb{R} \). Consider the line search map \( M : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \) defined by

\[ M(x, d) = \{ y : y = x + \lambda d \text{ for some } \lambda \in L \text{ and } f(y) \leq f(x + \lambda d) \text{ for each } \lambda \in L \}. \]

If \( f \) is continuous at \( x \) and \( d \neq 0 \), then \( M \) is closed at \( (x, d) \).

**Proof**

Suppose that \( (x_k, d_k) \to (x, d) \) and that \( y_k \to y \), where \( y_k \in M(x_k, d_k) \). We want to show that \( y \in M(x, d) \). First, note that \( y_k = x_k + \lambda_k d_k \), where \( \lambda_k \in L \). Since \( d \neq 0 \), \( d_k \neq 0 \) for \( k \) large enough, and hence \( \lambda_k = \| y_k - x_k \| / \| d_k \| \).

Taking the limit as \( k \to \infty \), then \( \lambda_k \to \lambda \), where \( \lambda = \| y - x \| / \| d \| \), and hence, \( y = x + \lambda d \). Furthermore, since \( \lambda_k \in L \) for each \( k \), and since \( L \) is closed, \( \lambda \in L \).

Now let \( \lambda \in L \) and note that \( f(y_k) \leq f(x_k + \lambda d_k) \) for all \( k \). Taking the limit as \( k \to \infty \) and noting the continuity of \( f \), we conclude that \( f(y) \leq f(x + \lambda d) \). Thus, \( y \in M(x, d) \), and the proof is complete.

In nonlinear programming, line search is typically performed over one of the following intervals:

\[ L = \{ \lambda : \lambda \in \mathbb{R} \} \]

\[ L = \{ \lambda : \lambda \geq 0 \} \]

\[ L = \{ \lambda : a \leq \lambda \leq b \}. \]

In each of the above cases, \( L \) is closed and the theorem applies.

In Theorem 8.4.1 we required that the vector \( d \) be nonzero. Example 8.4.2 presents a case in which \( M \) is not closed if \( d = 0 \). In most cases the direction vector \( d \) is nonzero over points outside the solution set \( \Omega \). Thus, \( M \) is closed at these points, and Theorem 7.2.3 can be applied to prove convergence.

8.4.2 Example

Consider the following problem:
Minimize \((x - 2)^4\).

Here \(f(x) = (x - 2)^4\). Now consider the sequence \((x_k, d_k) = (1/k, 1/k)\). Clearly, \(x_k\) converges to \(x = 0\) and \(d_k\) converges to \(d = 0\). Consider the line search map \(M\) defined in Theorem 8.4.1, where \(L = \{\lambda : \lambda \geq 0\}\). The point \(y_k\) is obtained by solving the problem to minimize \(f(x_k + \lambda d_k)\) subject to \(\lambda \geq 0\). The reader can verify that \(y_k = 2\) for all \(k\), so its limit \(y\) equals 2. Note, however, that \(M(0, 0) = \{0\}\), so that \(y \notin M(0, 0)\). This shows that \(M\) is not closed.

8.5 Multidimensional Search Without Using Derivatives

In this section we consider the problem of minimizing a function \(f\) of several variables without using derivatives. The methods described here proceed in the following manner. Given a vector \(x\), a suitable direction \(d\) is first determined, and then \(f\) is minimized from \(x\) in the direction \(d\) by one of the techniques discussed earlier in this chapter.

Throughout the book we are required to solve a line search problem of the form to minimize \(f(x + \lambda d)\) subject to \(\lambda \in L\), where \(L\) is typically of the form \(L = R, L = \{\lambda : \lambda \geq 0\}\) or \(L = \{\lambda : a \leq \lambda \leq b\}\). In the statements of the algorithms, for the purpose of simplicity we have assumed that a minimizing point \(\bar{x}\) exists. However, this may not be the case. Here, the optimal objective value of the line search problem may be unbounded, or else the optimal objective value may be finite but not achieved at any particular \(\lambda\). In the first case, the original problem is unbounded and we may stop. In the latter case, \(\lambda\) could be chosen as \(\bar{x}\) such that \(f(x + \bar{x}d)\) is sufficiently close to the value inf\{\(f(x + \lambda d) : \lambda \in L\}\}.

Cyclic Coordinate Method

This method uses the coordinate axes as the search directions. More specifically, the method searches along the directions \(d_1, \ldots, d_n\), where \(d_j\) is a vector of zeros except for a 1 at the \(j\)th position. Thus, along the search direction \(d_j\), the variable \(x_j\) is changed while all other variables are kept fixed. The method is illustrated schematically in Figure 8.7 for the problem of Example 8.5.1.

Note that we are assuming here that the minimization is done in order over the dimensions \(1, \ldots, n\) at each iteration. In a variant known as the Aitken double sweep method, the search is conducted by minimizing over the dimensions \(1, \ldots, n\) and then back over the dimensions \(n - 1, n - 2, \ldots, 1\). This requires \(n - 1\) line searches per iteration. Accordingly, if the function to be minimized is differentiable and its gradient is available, the Gauss-Southwell variant recommends that one select that coordinate direction for minimizing at each step that has the largest magnitude of the partial derivative component. These types of sequential one-dimensional minimizations are sometimes
referred to as *Gauss-Seidel iterations*, based on the *Gauss-Seidel method* for solving systems of equations.

**Summary of the Cyclic Coordinate Method**

We summarize below the cyclic coordinate method for minimizing a function of several variables without using any derivative information. As we show shortly, if the function is differentiable, the method converges to a stationary point.

As discussed in Section 7.2, several criteria could be used for terminating the algorithm. In the statement of the algorithm below, the termination criterion \( \|x_{k+1} - x_k\| < \varepsilon \) is used. Obviously, any of the other criteria could be used to stop the procedure.

**Initialization Step** Choose a scalar \( \varepsilon > 0 \) to be used for terminating the algorithm, and let \( d_1, \ldots, d_n \) be the coordinate directions. Choose an initial point \( x_1 \), let \( y_1 = x_1 \), let \( k = j = 1 \), and go to the Main Step.

**Main Step**

1. Let \( \lambda_j \) be an optimal solution to the problem to minimize \( f(y_j + \lambda d_j) \) subject to \( \lambda \in \mathbb{R} \), and let \( y_{j+1} = y_j + \lambda_j d_j \). If \( j < n \), replace \( j \) by \( j + 1 \), and repeat Step 1. Otherwise, if \( j = n \), go to Step 2.

2. Let \( x_{k+1} = y_{n+1} \). If \( \|x_{k+1} - x_k\| < \varepsilon \), then stop. Otherwise, let \( y_1 = x_{k+1} \), let \( j = 1 \), replace \( k \) by \( k + 1 \), and go to Step 1.

**8.5.1 Example**

Consider the following problem:

\[
\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.
\]

Note that the optimal solution to this problem is \((2, 1)\) with objective value equal to zero. Table 8.6 gives a summary of computations for the cyclic coordinate method starting from the initial point \((0, 3)\). Note that at each iteration, the vectors \( y_2 \) and \( y_3 \) are obtained by performing a line search in the directions \((1, 0)\) and \((0, 1)\), respectively. Also note that significant progress is made during the first few iterations, whereas much slower progress is made during later iterations. After seven iterations, the point \((2.22, 1.11)\), whose objective value is 0.0023, is reached.

In Figure 8.7 the contours of the objective function are given, and the points generated above by the cyclic coordinate method are shown. Note that at later iterations, slow progress is made because of the short orthogonal movements along the valley indicated by the dashed lines. Later, we analyze the convergence rate of steepest descent methods. The cyclic coordinate method tends to exhibit a performance characteristic over the \( n \) coordinate line searches similar to that of an iteration of the steepest descent method.
Table 8.6 Summary of Computations for the Cyclic Coordinate Method

<table>
<thead>
<tr>
<th>Iteration $k$</th>
<th>$x_k$</th>
<th>$f(x_k)$</th>
<th>$j$</th>
<th>$d_j$</th>
<th>$y_j$</th>
<th>$\lambda_j$</th>
<th>$y_{j+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.00, 3.00)</td>
<td>1</td>
<td>(1.0, 0.0)</td>
<td>(0.00, 3.00)</td>
<td>3.13</td>
<td>(3.13, 3.00)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>52.00</td>
<td>2</td>
<td>(0.0, 1.0)</td>
<td>(3.13, 3.00)</td>
<td>-1.44</td>
<td>(3.13, 1.56)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(3.13, 1.56)</td>
<td>1</td>
<td>(1.0, 0.0)</td>
<td>(3.13, 1.56)</td>
<td>-0.50</td>
<td>(2.63, 1.56)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.63</td>
<td>2</td>
<td>(0.0, 1.0)</td>
<td>(2.63, 1.56)</td>
<td>-0.25</td>
<td>(2.63, 1.31)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(2.63, 1.31)</td>
<td>1</td>
<td>(1.0, 0.0)</td>
<td>(2.63, 1.31)</td>
<td>-0.19</td>
<td>(2.44, 1.31)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>2</td>
<td>(0.0, 1.0)</td>
<td>(2.44, 1.31)</td>
<td>-0.09</td>
<td>(2.44, 1.22)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(2.44, 1.22)</td>
<td>1</td>
<td>(1.0, 0.0)</td>
<td>(2.44, 1.22)</td>
<td>-0.09</td>
<td>(2.35, 1.22)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>2</td>
<td>(0.0, 1.0)</td>
<td>(2.35, 1.22)</td>
<td>-0.05</td>
<td>(2.35, 1.17)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>(2.35, 1.17)</td>
<td>1</td>
<td>(1.0, 0.0)</td>
<td>(2.35, 1.17)</td>
<td>-0.06</td>
<td>(2.29, 1.17)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.015</td>
<td>2</td>
<td>(0.0, 1.0)</td>
<td>(2.29, 1.17)</td>
<td>-0.03</td>
<td>(2.29, 1.14)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>(2.29, 1.14)</td>
<td>1</td>
<td>(1.0, 0.0)</td>
<td>(2.29, 1.14)</td>
<td>-0.04</td>
<td>(2.25, 1.14)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.007</td>
<td>2</td>
<td>(0.0, 1.0)</td>
<td>(2.25, 1.14)</td>
<td>-0.02</td>
<td>(2.25, 1.12)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>(2.25, 1.12)</td>
<td>1</td>
<td>(1.0, 0.0)</td>
<td>(2.22, 1.12)</td>
<td>-0.03</td>
<td>(2.22, 1.12)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.004</td>
<td>2</td>
<td>(0.0, 1.0)</td>
<td>(2.22, 1.12)</td>
<td>-0.01</td>
<td>(2.22, 1.11)</td>
<td></td>
</tr>
</tbody>
</table>

Convergence of the Cyclic Coordinate Method

Convergence of the cyclic coordinate method to a stationary point follows immediately from Theorem 7.3.5 under the following assumptions:

1. The minimum of $f$ along any line in $R^n$ is unique.

2. The sequence of points generated by the algorithm is contained in a compact subset of $R^n$.

Note that the search directions used at each iteration are the coordinate vectors, so that the matrix of search directions $D = I$. Obviously, Assumption 1 of Theorem 7.3.5 holds true.

As an alternative approach, Theorem 7.2.3 could have been used to prove convergence after showing that the overall algorithmic map is closed at each $x$ satisfying $\nabla f(x) \neq 0$. In this case, the descent function $a$ is taken as $f$ itself, and the solution set is $\Omega = \{x: \nabla f(x) = 0\}$.

Acceleration Step

We learned from the foregoing analysis that when applied to a differentiable function, the cyclic coordinate method will converge to a point with zero gradient. In the absence of differentiability, however, the method can stall at a nonoptimal point. As shown in Figure 8.8a, searching along any of the
Figure 8.7 Cyclic coordinate method.

Coordinate axes at the point \( x_2 \) leads to no improvement of the objective function and results in premature termination. The reason for this premature termination is the presence of a sharp-edged valley caused by the nondifferentiability of \( f \). As illustrated in Figure 8.8b, this difficulty could possibly be overcome by searching along the direction \( x_2 - x_1 \).

The search along a direction \( x_{k+1} - x_k \) is frequently used in applying the cyclic coordinate method, even in the case where \( f \) is differentiable. The usual rule of thumb is to apply it at every \( p \)th iteration. This modification to the cyclic coordinate method frequently accelerates convergence, particularly when the sequence of points generated zigzags along a valley. Such a step is usually referred to as an acceleration step or a pattern search step.

Method of Hooke and Jeeves

The method of Hooke and Jeeves performs two types of search: exploratory search and pattern search. The first two iterations of the procedure are illustrated in Figure 8.9. Given \( x_1 \), an exploratory search along the coordinate directions produces the point \( x_2 \). Now a pattern search along the direction \( x_2 - x_1 \) leads to
Figure 8.8 Effect of a sharp-edged valley.

the point $y$. Another exploratory search starting from $y$ gives the point $x_3$. The next pattern search is conducted along the direction $x_3 - x_2$, yielding $y'$. The process is then repeated.

Summary of the Method of Hooke and Jeeves Using Line Searches

As originally proposed by Hooke and Jeeves, the method does not perform any line search but rather takes discrete steps along the search directions, as we discuss later. Here we present a continuous version of the method using line searches along the coordinate directions $d_1, ..., d_n$ and the pattern direction.

Figure 8.9 Method of Hooke and Jeeves.
Chapter 8

Initialization Step  Choose a scalar $\varepsilon > 0$ to be used in terminating the algorithm. Choose a starting point $x_1$, let $y_1 = x_1$, let $k = j = 1$, and go to the Main Step.

Main Step

1. Let $\lambda_j$ be an optimal solution to the problem to minimize $f(y_j + \lambda d_j)$ subject to $\lambda \in R$, and let $y_{j+1} = y_j + \lambda_j d_j$. If $j < n$, replace $j$ by $j + 1$, and repeat Step 1. Otherwise, if $j = n$, let $x_{n+1} = y_{n+1}$. If $\|x_{k+1} - x_k\| < \varepsilon$, stop; otherwise, go to Step 2.

2. Let $d = x_{k+1} - x_k$, and let $\hat{\lambda}$ be an optimal solution to the problem to minimize $f(x_{k+1} + \lambda d)$ subject to $\lambda \in R$. Let $y_1 = x_{k+1} + \hat{\lambda} d$, let $j = 1$, replace $k$ by $k + 1$, and go to Step 1.

8.5.2 Example

Consider the following problem:

$$\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.$$ 

Note that the optimal solution is $(2.00, 1.00)$ with objective value equal to zero. Table 8.7 summarizes the computations for the method of Hooke and Jeeves, starting from the initial point $(0.00, 3.00)$. At each iteration, an exploratory search along the coordinate directions gives the points $y_2$ and $y_3$, and a pattern search along the direction $d = x_{k+1} - x_k$ gives the point $y_1$, except at iteration $k = 1$, where $y_1 = x_1$. Note that four iterations were required to move from the initial point to the optimal point $(2.00, 1.00)$ whose objective value is zero. At this point, $\|x_5 - x_4\| = 0.045$, and the procedure is terminated.

Figure 8.10 illustrates the points generated by the method of Hooke and Jeeves using line searches. Note that the pattern search has substantially improved the convergence behavior by moving along a direction that is almost parallel to the valley shown by dashed lines.

Convergence of the Method of Hooke and Jeeves

Suppose that $f$ is differentiable, and let the solution set $\Omega = \{\bar{x} : \nabla f(\bar{x}) = 0\}$.

Note that each iteration of the method of Hooke and Jeeves consists of an application of the cyclic coordinate method, in addition to a pattern search. Let the cyclic coordinate search be denoted by the map $B$ and the pattern search be denoted by the map $C$. Using an argument similar to that of Theorem 7.3.5, it follows that $B$ is closed. If the minimum of $f$ along any line is unique and letting $\alpha = f$, then
Table 8.7 Summary of Computations for the Method of Hooke and Jeeves Using Line Searches

<table>
<thead>
<tr>
<th>Iteration $k$</th>
<th>$x_k$</th>
<th>$y_j$</th>
<th>$d_j$</th>
<th>$\lambda_j$</th>
<th>$y_{j+1}$</th>
<th>$d$</th>
<th>$\hat{\lambda}$</th>
<th>$y_3 + \hat{\lambda}d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.00, 3.00)</td>
<td>1</td>
<td>(0.00, 3.00)</td>
<td>(1.0, 0.0)</td>
<td>3.13</td>
<td>(3.13, 3.00)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>(3.13, 3.00)</td>
<td>(0.0, 1.0)</td>
<td>-1.44</td>
<td>(3.13, 1.56)</td>
<td>(3.13, 1.44)</td>
<td>-0.10</td>
</tr>
<tr>
<td>2</td>
<td>(3.13, 1.56)</td>
<td>1</td>
<td>(2.82, 1.70)</td>
<td>(1.0, 0.0)</td>
<td>-0.12</td>
<td>(2.70, 1.70)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>(2.70, 1.70)</td>
<td>(0.0, 1.0)</td>
<td>-0.35</td>
<td>(2.70, 1.35)</td>
<td>(-0.43, -0.21)</td>
<td>1.50</td>
</tr>
<tr>
<td>3</td>
<td>(2.70, 1.35)</td>
<td>1</td>
<td>(2.06, 1.04)</td>
<td>(1.0, 0.0)</td>
<td>-0.02</td>
<td>(2.04, 1.04)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>(2.04, 1.04)</td>
<td>(0.0, 1.0)</td>
<td>-0.02</td>
<td>(2.04, 1.02)</td>
<td>(-0.66, -0.33)</td>
<td>0.06</td>
</tr>
<tr>
<td>4</td>
<td>(2.04, 1.02)</td>
<td>1</td>
<td>(2.00, 1.00)</td>
<td>(1.0, 0.0)</td>
<td>0.00</td>
<td>(2.00, 1.00)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>(2.00, 1.00)</td>
<td>(0.0, 1.0)</td>
<td>0.00</td>
<td>(2.00, 1.00)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>5</td>
<td>(2.00, 1.00)</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
<td>—</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>
Figure 8.10 Method of Hooke and Jeeves using line searches. Method of
Hooke and Jeeves with Discrete Steps

\( \alpha(y) < \alpha(x) \) for \( x \in \Omega \). By the definition of \( C \), \( \alpha(z) \leq \alpha(y) \) for \( z \in C(y) \).
Assuming that \( \Lambda = \{x : f(x) \leq f(x_1)\} \), where \( x_1 \) is the starting point, is compact, convergence of the procedure is established by Theorem 7.3.4.

**Method of Hooke and Jeeves with Discrete Steps**

As mentioned earlier, the method of Hooke and Jeeves, as originally proposed, does not perform line searches but, instead, adopts a simple scheme involving functional evaluations. A summary of the method is given below.

**Initialization Step**

Let \( d_1, \ldots, d_n \) be the coordinate directions. Choose a scalar \( \varepsilon > 0 \) to be used for terminating the algorithm. Furthermore, choose an initial step size, \( \Delta \geq \varepsilon \), and an acceleration factor, \( \alpha > 0 \). Choose a starting point \( x_1 \), let \( y_1 = x_1 \), let \( k = j = 1 \), and go to the Main Step.

**Main Step**

1. If \( f(y_j + \Delta d_j) < f(y_j) \), the trial is termed a success; let \( y_{j+1} = y_j + \Delta d_j \), and go to Step 2. If, however, \( f(y_j + \Delta d_j) \geq f(y_j) \), the
trial is deemed a failure. In this case, if \( f(y_j - \Delta d_j) < f(y_j) \), let \( y_{j+1} = y_j - \Delta d_j \), and go to Step 2; if \( f(y_j - \Delta d_j) \geq f(y_j) \), let \( y_{j+1} = y_j \), and go to Step 2.

2. If \( j < n \), replace \( j \) by \( j + 1 \), and repeat Step 1. Otherwise, go to Step 3 if \( f(y_{n+1}) < f(x_k) \), and go to Step 4 if \( f(y_{n+1}) \geq f(x_k) \).

3. Let \( x_{k+1} = y_{n+1} \), and let \( y_1 = x_{k+1} + \alpha(x_{k+1} - x_k) \). Replace \( k \) by \( k + 1 \), let \( j = 1 \), and go to Step 1.

4. If \( \Delta \leq \varepsilon \), stop; \( x_k \) is the prescribed solution. Otherwise, replace \( \Delta \) by \( \Delta/2 \).

Let \( y_1 = x_k \), \( x_{k+1} = x_k \), replace \( k \) by \( k + 1 \), let \( j = 1 \), and repeat Step 1.

The reader may note that steps 1 and 2 above describe an exploratory search. Furthermore, Step 3 is an acceleration step along the direction \( x_{k+1} - x_k \). Note that a decision whether to accept or reject the acceleration step is not made until after an exploratory search is performed. In Step 4, the step size \( \Delta \) is reduced. The procedure could easily be modified so that different step sizes are used along the different directions. This is sometimes adopted for the purpose of scaling.

### 8.5.3 Example

Consider the following problem:

\[
\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.
\]

We solve the problem using the method of Hooke and Jeeves with discrete steps. The parameters \( \alpha \) and \( \Delta \) are chosen as 1.0 and 0.2, respectively. Figure 8.11 shows the path taken by the algorithm starting from \((0.0, 3.0)\). The points generated are numbered sequentially, and the acceleration step that is rejected is shown by the dashed lines. From this particular starting point, the optimal solution is easily reached.

To give a more comprehensive illustration, Table 8.8 summarizes the computations starting from the new initial point \((2.0, 3.0)\). Here (S) denotes that the trial is a success and (F) denotes that the trial is a failure. At the first iteration, and at subsequent iterations whenever \( f(y_3) \geq f(x_k) \), the vector \( y_1 \) is taken as \( x_k \). Otherwise, \( y_1 = 2x_{k+1} - x_k \). Note that at the end of iteration \( k = 10 \), the point \((1.70, 0.80)\) is reached having an objective value 0.02. The procedure is stopped here with the termination parameter \( \varepsilon = 0.1 \). If a greater degree of accuracy is required, \( \Delta \) should be reduced to 0.05.

Figure 8.12 illustrates the path taken by the method. The points generated are again numbered sequentially, and dashed lines represent rejected acceleration steps.
Table 8.8 Summary of Computations for the Method of Hooke and Jeeves with Discrete Steps

<table>
<thead>
<tr>
<th>Iteration $k$</th>
<th>$\Delta$</th>
<th>$x_k$</th>
<th>$y_j$</th>
<th>$d_j$</th>
<th>$y_j + \Delta d_j$</th>
<th>$y_j - \Delta d_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$f(x_k)$</td>
<td>$f(y_j)$</td>
<td>$f(y_j + \Delta d_j)$</td>
<td>$f(y_j - \Delta d_j)$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.2</td>
<td>(2.00, 3.00)</td>
<td>16.00</td>
<td>16.00</td>
<td>(2.20, 3.00)</td>
<td>16.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2.00, 3.00)</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>(2.20, 3.00)</td>
<td>14.44</td>
<td>(0.0, 1.0)</td>
<td>(2.20, 2.80)</td>
<td>11.56</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
<td>(2.40, 2.60)</td>
<td>7.87</td>
<td>(1.0, 0.0)</td>
<td>(2.60, 2.60)</td>
<td>6.89</td>
</tr>
<tr>
<td>4</td>
<td>0.2</td>
<td>(2.60, 2.40)</td>
<td>1.85</td>
<td>(0.0, 1.0)</td>
<td>(2.80, 2.20)</td>
<td>2.97</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>(2.80, 1.80)</td>
<td>0.57</td>
<td>(0.0, 1.0)</td>
<td>(2.80, 1.40)</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2.80, 2.00)</td>
<td>1.36</td>
<td>(1.0, 0.0)</td>
<td>(2.80, 1.20)</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2.80, 1.00)</td>
<td>1.05</td>
<td>(1.0, 0.0)</td>
<td>(2.80, 1.00)</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2.60, 1.00)</td>
<td>0.49</td>
<td>(0.0, 1.0)</td>
<td>(2.60, 1.20)</td>
<td>0.17</td>
</tr>
</tbody>
</table>

(continued)
<p>| | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.2</td>
<td>(2.60, 1.20)</td>
<td>1</td>
<td>(2.40, 1.00)</td>
<td>(1.0, 0.0)</td>
<td>(2.60, 1.00)</td>
<td>(2.20, 1.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.17</td>
<td></td>
<td>0.19</td>
<td></td>
<td>0.49(F)</td>
<td></td>
<td>0.04(S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>(2.20, 1.00)</td>
<td>(0.0, 1.0)</td>
<td>(2.20, 1.20)</td>
<td>(2.20, 0.80)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.04</td>
<td></td>
<td>0.04(F)</td>
<td></td>
<td>0.36(F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.2</td>
<td>(2.20, 1.00)</td>
<td>1</td>
<td>(1.80, 0.80)</td>
<td>(1.0, 0.0)</td>
<td>(2.00, 0.80)</td>
<td>(1.60, 0.80)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.04</td>
<td></td>
<td>0.04</td>
<td></td>
<td>0.16(F)</td>
<td></td>
<td>0.03(S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>(1.60, 0.80)</td>
<td>(0.0, 1.0)</td>
<td>(1.60, 1.00)</td>
<td>(1.60, 0.60)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.03</td>
<td></td>
<td>0.19(F)</td>
<td></td>
<td>0.19(F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.2</td>
<td>(1.60, 0.80)</td>
<td>1</td>
<td>(1.00, 0.60)</td>
<td>(1.0, 0.0)</td>
<td>(1.20, 0.60)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.03</td>
<td></td>
<td>0.67</td>
<td></td>
<td>0.41(S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>(1.20, 0.60)</td>
<td>(0.0, 1.0)</td>
<td>(1.20, 0.80)</td>
<td>(1.20, 0.40)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.41</td>
<td></td>
<td>0.57(F)</td>
<td></td>
<td>0.57(F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.1</td>
<td>(1.60, 0.80)</td>
<td>1</td>
<td>(1.60, 0.80)</td>
<td>(1.0, 0.0)</td>
<td>(1.70, 0.80)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.03</td>
<td></td>
<td>0.03</td>
<td></td>
<td>0.02(S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>(1.70, 0.80)</td>
<td>(0.0, 1.0)</td>
<td>(1.70, 0.90)</td>
<td>(1.70, 0.70)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.02</td>
<td></td>
<td>0.02(F)</td>
<td></td>
<td>0.10(F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.1</td>
<td>(1.70, 0.80)</td>
<td>1</td>
<td>(1.80, 0.80)</td>
<td>(1.0, 0.0)</td>
<td>(1.90, 0.80)</td>
<td>(1.70, 0.80)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.02</td>
<td></td>
<td>0.04</td>
<td></td>
<td>0.09(F)</td>
<td></td>
<td>0.02(S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>(1.70, 0.80)</td>
<td>(0.0, 1.0)</td>
<td>(1.70, 0.90)</td>
<td>(1.70, 0.70)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.02</td>
<td></td>
<td>0.02(F)</td>
<td></td>
<td>0.10(F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Method of Rosenbrock

As originally proposed, the method of Rosenbrock does not employ line searches but rather takes discrete steps along the search directions. We present here a continuous version of the method that utilizes line searches. At each iteration, the procedure searches iteratively along $n$ linearly independent and orthogonal directions. When a new point is reached at the end of an iteration, a new set of orthogonal vectors is constructed. In Figure 8.13 the new directions are denoted by $d_1$ and $d_2$.

Construction of the Search Directions

Let $d_1, \ldots, d_n$ be linearly independent vectors, each with a norm equal to 1. Furthermore, suppose that these vectors are mutually orthogonal; that is, $d_i^td_j = 0$ for $i \neq j$. Starting from the current vector $x_k$, the objective function $f$ is
minimized along each of the directions iteratively, resulting in the point $x_{k+1}$. In particular, $x_{k+1} - x_k = \sum_{j=1}^{n} \lambda_j d_j$, where $\lambda_j$ is the distance moved along $d_j$.

The new collection of directions $\overrightarrow{d}_1, ..., \overrightarrow{d}_n$ is formed by the Gram-Schmidt procedure, or orthogonalization procedure, as follows:

$$a_j = \begin{cases} 
  d_j & \text{if } \lambda_j = 0 \\
  \sum_{i=j}^{n} \lambda_i d_i & \text{if } \lambda_j \neq 0 
\end{cases}$$

$$b_j = \begin{cases} 
  a_j, & j = 1 \\
  a_j - \sum_{i=1}^{j} (a_i' \overrightarrow{d}_i) \overrightarrow{d}_i, & j \geq 2 
\end{cases}$$

$$\overrightarrow{d}_j = \frac{b_j}{\|b_j\|}.$$
Lemma 8.5.4 shows that the new directions established by the Rosenbrock procedure are indeed linearly independent and orthogonal.

8.5.4 Lemma
Suppose that the vectors $d_1, \ldots, d_n$ are linearly independent and mutually orthogonal. Then the directions $\overline{d}_1, \ldots, \overline{d}_n$ defined by (8.9) are also linearly independent and mutually orthogonal for any set of $\lambda_1, \ldots, \lambda_n$. Furthermore, if $\lambda_j = 0$, then $\overline{d}_j = d_j$.

Proof
We first show that $a_1, \ldots, a_n$ are linearly independent. Suppose that $\sum_{j=1}^{n} \mu_j a_j = 0$. Let $I = \{j : \lambda_j = 0\}$, and let $J(j) = \{i : i \not\in I, i \leq j\}$. Noting (8.9), we get

$$0 = \sum_{j=1}^{n} \mu_j a_j = \sum_{j \in I} \mu_j d_j + \sum_{j \not\in I} \mu_j \left( \sum_{i=j}^{n} \lambda_i d_i \right)$$

$$= \sum_{j \in I} \mu_j d_j + \sum_{j \not\in I} \left( \lambda_j \sum_{i \in J(j)} \mu_i \right) d_j.$$

Since $d_1, \ldots, d_n$ are linearly independent, $\mu_j = 0$ for $j \in I$ and $\lambda_j \sum_{i \in J(j)} \mu_i = 0$ for $j \not\in I$. But $\lambda_j \neq 0$ for $j \not\in I$, and hence, $\sum_{i \in J(j)} \mu_i = 0$ for each $j \not\in I$. By the definition of $J(j)$, we therefore have $\mu_1 = \cdots = \mu_n = 0$, and hence, $a_1, \ldots, a_n$ are linearly independent.
To show that $b_1, \ldots, b_n$ are linearly independent, we use the following induction argument. Since $b_1 = a_1 \neq 0$, it suffices to show that if $b_1, \ldots, b_k$ are linearly independent, then $b_1, \ldots, b_k, b_{k+1}$ are also linearly independent. Suppose that $\sum_{j=1}^{k+1} \alpha_j b_j = 0$. Using the definition of $b_{k+1}$, in (8.9), we get

$$0 = \sum_{j=1}^{k} \alpha_j b_j + \alpha_{k+1} b_{k+1}$$

$$= \sum_{j=1}^{k} \frac{\alpha_{k+1} (a_{k+1}^t d_j)}{\|b_j\|} b_j + \alpha_{k+1} a_{k+1}.$$  \hspace{1cm} (8.10)

From (8.9) it follows that each vector $b_j$ is a linear combination of $a_1, \ldots, a_j$. Since $a_1, \ldots, a_{k+1}$ are linearly independent, it follows from (8.10) that $\alpha_{k+1} = 0$. Since $b_1, \ldots, b_k$ are assumed linearly independent by the induction hypotheses, from (8.10) we get $\alpha_j = \alpha_{k+1} (a_{k+1}^t \bar{d}_j) / \|b_j\| = 0$ for $j = 1, \ldots, k$. Since $\alpha_{k+1} = 0$, $\alpha_j = 0$ for each $j$. This shows that $b_1, \ldots, b_{k+1}$ are linearly independent. By the definition of $d_j$, linear independence of $d_1, \ldots, d_n$ is immediate.

Now we establish the orthogonality of $b_1, \ldots, b_n$ and hence the orthogonality of $d_1, \ldots, d_n$. From (8.9), $b_1^t b_2 = 0$; thus, it suffices to show that if $b_1, \ldots, b_k$ are mutually orthogonal, then $b_1, \ldots, b_k, b_{k+1}$ are also mutually orthogonal. From (8.10) and noting that $b_j^t d_i = 0$ for $i \neq j$, it follows that

$$b_j^t b_{k+1} = b_j^t \left[ a_{k+1} - \sum_{i=1}^{k} (a_{k+1}^t \bar{d}_i) \bar{d}_i \right]$$

$$= b_j^t a_{k+1} - (a_{k+1}^t d_j) b_j^t d_j = 0.$$  \hspace{1cm}

Thus, $b_1, \ldots, b_{k+1}$ are mutually orthogonal.

To complete the proof, we show that $\bar{d}_j = d_j$ if $\lambda_j = 0$. From (8.9), if $\lambda_j = 0$, we get

$$b_j = d_j - \sum_{i=1}^{j-1} \frac{1}{\|b_i\|} (d_i^t b_i) \bar{d}_i.$$  \hspace{1cm} (8.11)

Note that $b_i$ is a linear combination of $a_1, \ldots, a_j$, so that $b_i = \sum_{r=1}^{i} \beta_{ir} a_r$. From (8.9), it thus follows that
\[ b_i = \sum_{r \in \mathcal{R}} \beta_{ir} d_r + \sum_{r \in \mathcal{R}} \beta_{ir} \left( \sum_{s=r}^{n} \lambda_s d_s \right), \quad (8.12) \]

where \( \mathcal{R} = \{ r : r \leq i, \lambda_r = 0 \} \) and \( \mathcal{R} = \{ r : r \leq i, \lambda_r \neq 0 \} \). Consider \( i < j \) and note that \( d'_j d_v = 0 \) for \( v \neq j \). For \( r \in \mathcal{R}, r \leq i < j \) and hence \( d'_j d_r = 0 \). For \( r \not\in \mathcal{R}, d'_j (\sum_{s=r}^{n} \lambda_s d_s) = \lambda_j d'_j d_j = \lambda_j \). By assumption, \( \lambda_j = 0 \), and thus multiplying \((8.12)\) by \( d'_j \), we get \( d'_j b_j = 0 \) for \( i < j \). From \((8.11)\) it follows that \( b_j = d_j \), and hence, \( \bar{d}_j = d_j \). This completes the proof.

From Lemma 8.5.4, if \( \lambda_j = 0 \), then the new direction \( \bar{d}_j \) is equal to the old direction \( d_j \). Hence, we only need to compute new directions for those indices with \( \lambda_j \neq 0 \).

**Summary of the Method of Rosenbrock Using Line Searches**

We now summarize Rosenbrock’s method using line searches for minimizing a function \( f \) of several variables. As we shall show shortly, if \( f \) is differentiable, then the method converges to a point with zero gradient.

**Initialization Step**  Let \( \varepsilon > 0 \) be the termination scalar. Choose \( d_1, \ldots, d_n \) as the coordinate directions. Choose a starting point \( x_1 \), let \( y_1 = x_1 \), \( k = j = 1 \), and go to the Main Step.

**Main Step**

1. Let \( \lambda_j \) be an optimal solution to the problem to minimize \( f(y_j + \lambda d_j) \) subject to \( \lambda \in \mathcal{R} \), and let \( y_{j+1} = y_j + \lambda_j d_j \). If \( j < n \), replace \( j \) by \( j + 1 \), and repeat Step 1. Otherwise, go to Step 2.

2. Let \( x_{k+1} = y_{n+1} \). If \( \| x_{k+1} - x_k \| < \varepsilon \) then stop; otherwise, let \( y_1 = x_{k+1} \), replace \( k \) by \( k + 1 \), let \( j = 1 \), and go to Step 3.

3. Form a new set of linearly independent orthogonal search directions by \((8.9)\). Denote these new directions by \( d_1, \ldots, d_n \) and go to Step 1.

**8.5.5 Example**

Consider the following problem:

\[ \text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2. \]

We solve this problem by the method of Rosenbrock using line searches. Table 8.9 summarizes the computations starting from the point \((0.00, 3.00)\). The
Table 8.9 Summary of Computations for the Method of Rosenbrock Using Line Searches

<table>
<thead>
<tr>
<th>Iteration $k$</th>
<th>$x_k$</th>
<th>$y_j$</th>
<th>$d_j$</th>
<th>$\lambda_j$</th>
<th>$y_{j+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x_k)$</td>
<td>$f(y_j)$</td>
<td>$f(y_{j+1})$</td>
<td>$f(y_{j+1})$</td>
<td>$f(y_{j+1})$</td>
<td>$f(y_{j+1})$</td>
</tr>
<tr>
<td>1</td>
<td>(0.00, 3.00)</td>
<td>52.00</td>
<td>1 (0.00, 3.00)</td>
<td>(1.00, 0.00)</td>
<td>3.13</td>
</tr>
<tr>
<td>2</td>
<td>(3.13, 3.00)</td>
<td>9.87</td>
<td>2 (3.13, 0.00)</td>
<td>(0.00, 1.00)</td>
<td>-1.44</td>
</tr>
<tr>
<td>3</td>
<td>(3.13, 1.56)</td>
<td>1.63</td>
<td>3 (3.13, 1.56)</td>
<td>(0.91, -0.42)</td>
<td>-0.34</td>
</tr>
<tr>
<td>4</td>
<td>(2.61, 1.24)</td>
<td>0.16</td>
<td>4 (2.61, 1.24)</td>
<td>(-0.85, -0.52)</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>(2.29, 1.04)</td>
<td>0.05</td>
<td>5 (2.29, 1.04)</td>
<td>(0.52, -0.85)</td>
<td>-0.10</td>
</tr>
<tr>
<td>6</td>
<td>(2.24, 1.13)</td>
<td>0.004</td>
<td>6 (2.24, 1.13)</td>
<td>(-0.96, -0.28)</td>
<td>0.04</td>
</tr>
<tr>
<td>7</td>
<td>(2.20, 1.12)</td>
<td>0.003</td>
<td>7 (2.20, 1.12)</td>
<td>(0.28, -0.96)</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Point $y_2$ is obtained by optimizing the function along the direction $d_1$ starting from $y_1$, and $y_3$ is obtained by optimizing the function along the direction $d_2$ starting from $y_2$. After the first iteration, we have $\lambda_1 = 3.13$ and $\lambda_2 = -1.44$. Using (8.9), the new search directions are $(0.91, -0.42)$ and $(-0.42, -0.91)$. After four iterations, the point $(2.21, 1.10)$ is reached, and the corresponding objective function value is 0.002. We now have $\|x_4 - x_3\| = 0.15$, and the procedure is stopped.

In Figure 8.14 the progress of the method is shown. It may be interesting to compare this figure with Figure 8.15, which is given later for the method of Rosenbrock using discrete steps.

**Convergence of the Method of Rosenbrock**

Note that according to Lemma 8.5.4, the search directions employed by the method are linearly independent and mutually orthogonal, and each has norm 1. Thus, at any given iteration, the matrix $D$ denoting the search directions satisfies $D^TD = I$. Thus, $\det[D] = 1$ and hence Assumption 1 of Theorem 7.3.5 holds true. By this theorem it follows that the method of Rosenbrock using line searches converges to a stationary point if the following assumptions are true:

Unconstrained Optimization
1. The minimum of $f$ along any line in $\mathbb{R}^n$ is unique.
2. The sequence of points generated by the algorithm is contained in a compact subset of $\mathbb{R}^n$.

Rosenbrock’s Method with Discrete Steps

As mentioned earlier, the method proposed by Rosenbrock avoids line searches. Instead, functional values are made at specific points. Furthermore, an acceleration feature is incorporated by suitably increasing or decreasing the step lengths as the method proceeds. A summary of the method is given below.

Initialization Step  Let $\varepsilon > 0$ be the termination scalar, let $\alpha > 1$ be a chosen expansion factor, and let $\beta \in (-1, 0)$ be a selected contraction factor. Choose $d_1, \ldots, d_n$ as the coordinate directions, and let $\Delta_1, \ldots, \Delta_n > 0$ be the initial step sizes along these directions. Choose a starting point $x_1$, let $y_1 = x_1$, $k = j = 1$, let $\Delta_j = \Delta_j$ for each $j$, and go to the Main Step.

Main Step
1. If $f(y_j + \Delta_j d_j) < f(y_j)$, the $j$th trial is deemed a success; set $y_{j+1} = y_j + \Delta_j d_j$, and replace $\Delta_j$ by $\alpha \Delta_j$. If, on the other hand, $f(y_j + \Delta_j d_j) \geq f(y_j)$, the trial is considered a failure; set $y_{j+1} = y_j$, and...
replace $\Delta_j$ by $\beta\Delta_j$. If $j < n$, replace $j$ by $j + 1$, and repeat Step 1. Otherwise, if $j = n$, go to Step 2.

2. If $f(y_{n+1}) < f(y_1)$, that is, if any of the $n$ trials of Step 1 were successful, let $y_1 = y_{n+1}$, set $j = 1$, and repeat Step 1. Now consider the case $f(y_{n+1}) = f(y_1)$, that is, when each of the last $n$ trials of Step 1 was a failure. If $f(y_{n+1}) < f(x_k)$, that is, if at least one successful trial was encountered in Step 1 during iteration $k$, go to Step 3. If $f(y_{n+1}) = f(x_k)$, that is, if no successful trial is encountered, stop with $x_k$ as an estimate of the optimal solution if $|\Delta_j| \leq \varepsilon$ for $j$; otherwise, let $y_1 = y_{n+1}$, let $j = 1$, and go to Step 1.

3. Let $x_{k+1} = y_{n+1}$. If $\|x_{k+1} - x_k\| < \varepsilon$, stop with $x_{k+1}$ as an estimate of the optimal solution. Otherwise, compute $\lambda_1, \ldots, \lambda_n$ from the relationship $x_{k+1} - x_k = \sum_{j=1}^n \lambda_j d_j$, form a new set of search directions by (8.9) and denote these directions by $d_1, \ldots, d_n$, let $\Delta_j = \bar{\Delta}_j$ for each $j$, let $y_1 = x_{k+1}$, replace $k$ by $k + 1$, let $j = 1$, and go to Step 1.

Note that discrete steps are taken along the $n$ search directions in Step 1. If a success occurs along $d_j$, then $\Delta_j$ is replaced by $\alpha\Delta_j$; and if a failure occurs along $d_j$, then $\Delta_j$ is replaced by $\beta\Delta_j$. Since $\beta < 0$, a failure results in reversing the $j$th search direction during the next pass through Step 1. Note that Step 1 is repeated until a failure occurs along each of the search directions, in which case, if at least one success was obtained during a previous loop at this iteration, a new set of search directions is formed by the Gram–Schmidt procedure. If the loops through the search directions continue to result in failures, the step length shrinks to zero.

8.5.6 Example

Consider the following problem:

$$
\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.
$$

We solve this problem by the method of Rosenbrock using discrete steps with $\bar{\Delta}_1 = \bar{\Delta}_2 = 0.1$, $\alpha = 2.0$, and $\beta = -0.5$. Table 8.10 summarizes the computations starting from $(0.00, 3.00)$, where (S) denotes a success and (F) denotes a failure. Note that within each iteration the directions $d_1$ and $d_2$ are fixed. After seven passes through Step 1 of Rosenbrock's method, we move from $x_1' = (0.00, 3.00)$ to $x_2' = (3.10, 1.45)$. At this point, a change of directions is required. In particular, $(x_2 - x_1) = \lambda_1 d_1 + \lambda_2 d_2$, where $\lambda_1 = 3.10$ and $\lambda_2 = -1.55$. Using (8.9), the reader can easily verify that the new search directions are
given by $(0.89, -0.45)$ and $(-0.45, -0.89)$, which are used in the second iteration. The procedure is terminated during the second iteration.

Figure 8.15 displays the progress of Rosenbrock’s method, where the points generated are numbered sequentially.

### 8.6 Multidimensional Search Using Derivatives

In the preceding section we described several minimization procedures that use only functional evaluations during the course of optimization. We now discuss some methods that use derivatives in determining the search directions. In particular, we discuss the steepest descent method and the method of Newton.

#### Method of Steepest Descent

The method of steepest descent, proposed by Cauchy in 1847, is one of the most fundamental procedures for minimizing a differentiable function of several variables. Recall that a vector $d$ is called a direction of descent of a function $f$ at $x$ if there exists a $\delta > 0$ such that $f(x + \lambda d) < f(x)$ for all $\lambda \in (0, \delta)$. In particular, if $\lim_{\lambda \to 0^+} [f(x + \lambda d) - f(x)]/\lambda < 0$, then $d$ is a direction of descent. The method of steepest descent moves along the direction $d$ with $\|d\| = 1$, which minimizes the above limit. Lemma 8.6.1 shows that if $f$ is differentiable at $x$ with a nonzero gradient, then $-\nabla f(x)/\|\nabla f(x)\|$ is indeed the direction of steepest descent. For this reason, in the presence of differentiability, the method of steepest descent is sometimes called the gradient method; it is also referred to as Cauchy’s method.

![Figure 8.15 Rosenbrock’s procedure using discrete steps. (The numbers denote the order in which points are generated.)](image)
8.6.1 Lemma

Suppose that \( f: \mathbb{R}^n \to \mathbb{R} \) is differentiable at \( x \), and suppose that \( \nabla f(x) \neq 0 \). Then the optimal solution to the problem to minimize \( f(x; d) \) subject to \( \|d\| \leq 1 \) is given by \( \tilde{d} = -\nabla f(x)/\|\nabla f(x)\| \); that is, \( -\nabla f(x)/\|\nabla f(x)\| \) is the direction of steepest descent of \( f \) at \( x \).

<table>
<thead>
<tr>
<th>Iteration</th>
<th>( x_k )</th>
<th>( y_j )</th>
<th>( \Delta_j )</th>
<th>( d_j )</th>
<th>( y_j + \Delta_j d_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>((0.00, 3.00))</td>
<td>((0.00, 3.00))</td>
<td>0.10</td>
<td>((1.00, 0.00))</td>
<td>((0.10, 3.00))</td>
</tr>
<tr>
<td></td>
<td>52.00</td>
<td>52.00</td>
<td></td>
<td></td>
<td>47.84(S)</td>
</tr>
<tr>
<td>2</td>
<td>((0.10, 3.00))</td>
<td>((0.00, 1.00))</td>
<td>0.10</td>
<td>((0.00, 1.00))</td>
<td>((0.10, 3.10))</td>
</tr>
<tr>
<td></td>
<td>47.84</td>
<td></td>
<td></td>
<td></td>
<td>50.24(F)</td>
</tr>
<tr>
<td>1</td>
<td>((0.10, 3.00))</td>
<td>((1.00, 0.00))</td>
<td>0.20</td>
<td>((1.00, 0.00))</td>
<td>((0.30, 3.00))</td>
</tr>
<tr>
<td></td>
<td>47.84</td>
<td></td>
<td></td>
<td></td>
<td>40.84(S)</td>
</tr>
<tr>
<td>2</td>
<td>((0.30, 3.00))</td>
<td>((0.00, 1.00))</td>
<td>-0.05</td>
<td>((0.00, 1.00))</td>
<td>((0.30, 2.95))</td>
</tr>
<tr>
<td></td>
<td>40.84</td>
<td></td>
<td></td>
<td></td>
<td>39.71(S)</td>
</tr>
<tr>
<td>1</td>
<td>((0.30, 2.95))</td>
<td>((1.00, 0.00))</td>
<td>0.40</td>
<td>((1.00, 0.00))</td>
<td>((0.70, 2.95))</td>
</tr>
<tr>
<td></td>
<td>39.71</td>
<td></td>
<td></td>
<td></td>
<td>29.90(S)</td>
</tr>
<tr>
<td>2</td>
<td>((0.70, 2.95))</td>
<td>((0.00, 1.00))</td>
<td>-0.10</td>
<td>((0.00, 1.00))</td>
<td>((0.70, 2.85))</td>
</tr>
<tr>
<td></td>
<td>29.90</td>
<td></td>
<td></td>
<td></td>
<td>27.86(S)</td>
</tr>
<tr>
<td>1</td>
<td>((0.70, 2.85))</td>
<td>((1.00, 0.00))</td>
<td>0.80</td>
<td>((1.00, 0.00))</td>
<td>((1.50, 2.85))</td>
</tr>
<tr>
<td></td>
<td>27.86</td>
<td></td>
<td></td>
<td></td>
<td>17.70(S)</td>
</tr>
<tr>
<td>2</td>
<td>((1.50, 2.85))</td>
<td>((0.00, 1.00))</td>
<td>-0.20</td>
<td>((0.00, 1.00))</td>
<td>((1.50, 2.65))</td>
</tr>
<tr>
<td></td>
<td>17.70</td>
<td></td>
<td></td>
<td></td>
<td>14.50(S)</td>
</tr>
<tr>
<td>1</td>
<td>((1.50, 2.65))</td>
<td>((1.00, 0.00))</td>
<td>1.60</td>
<td>((1.00, 0.00))</td>
<td>((3.10, 2.65))</td>
</tr>
<tr>
<td></td>
<td>14.50</td>
<td></td>
<td></td>
<td></td>
<td>6.30(S)</td>
</tr>
<tr>
<td>2</td>
<td>((3.10, 2.65))</td>
<td>((0.00, 1.00))</td>
<td>-0.40</td>
<td>((0.00, 1.00))</td>
<td>((3.10, 2.25))</td>
</tr>
<tr>
<td></td>
<td>6.30</td>
<td></td>
<td></td>
<td></td>
<td>3.42(S)</td>
</tr>
<tr>
<td>1</td>
<td>((3.10, 2.25))</td>
<td>((1.00, 0.00))</td>
<td>3.20</td>
<td>((1.00, 0.00))</td>
<td>((6.30, 2.25))</td>
</tr>
<tr>
<td></td>
<td>3.42</td>
<td></td>
<td></td>
<td></td>
<td>345.12(F)</td>
</tr>
<tr>
<td>2</td>
<td>((3.10, 2.25))</td>
<td>((0.00, 1.00))</td>
<td>-0.80</td>
<td>((0.00, 1.00))</td>
<td>((3.10, 1.45))</td>
</tr>
<tr>
<td></td>
<td>3.42</td>
<td></td>
<td></td>
<td></td>
<td>1.50(S)</td>
</tr>
</tbody>
</table>

(continued)
From the differentiability of $f$ at $x$, it follows that

$$f'(x; d) = \lim_{\lambda \to 0^+} \frac{f(x + \lambda d) - f(x)}{\lambda} = \nabla f(x)^T d.$$ 

Thus, the problem reduces to minimizing $\nabla f(x)^T d$ subject to $\|d\| \leq 1$. By the Schwartz inequality, for $\|d\| \leq 1$ we have

$$\nabla f(x)^T d \geq -\|\nabla f(x)\| \|d\| \geq -\|\nabla f(x)\|,$$

with equality holding throughout if and only if $d = \overline{d} = -\nabla f(x)/\|\nabla f(x)\|$. Thus, $\overline{d}$ is the optimal solution, and the proof is complete.
Summary of the Steepest Descent Algorithm

Given a point \( x \), the steepest descent algorithm proceeds by performing a line search along the direction \( -\nabla f(x)/\|\nabla f(x)\| \) or, equivalently, along the direction \( -\nabla f(x) \). A summary of the method is given below.

**Initialization Step**  Let \( \varepsilon > 0 \) be the termination scalar. Choose a starting point \( x_1 \), let \( k = 1 \), and go to the Main Step.

**Main Step**

If \( \|\nabla f(x_k)\| < \varepsilon \), stop; otherwise, let \( d_k = -\nabla f(x_k) \), and let \( \lambda_k \) be an optimal solution to the problem to minimize \( f(x_k + \lambda d_k) \) subject to \( \lambda \geq 0 \). Let \( x_{k+1} = x_k + \lambda_k d_k \), replace \( k \) by \( k + 1 \), and repeat the Main Step.

### 8.6.2 Example

Consider the following problem:

\[
\text{Minimize } (x_1 - 2)^4 + (x_1 - 2)^2.
\]

We solve this problem using the method of steepest descent, starting with the point \((0.00, 3.00)\). A summary of the computations is given in Table 8.11. After seven iterations, the point \( x_8 = (2.28, 1.15) \) is reached. The algorithm is terminated since \( \|\nabla f(x_8)\| = 0.09 \) is small. The progress of the method is shown in Figure 8.16. Note that the minimizing point for this problem is \((2.00, 1.00)\).

### Convergence of the Steepest Descent Method

Let \( \Omega = \{ \overline{x} : \nabla f(\overline{x}) = 0 \} \), and let \( f \) be the descent function. The algorithmic map is \( A = MD \), where \( D(x) = [x, \nabla f(x)] \) and \( M \) is the line search map over the closed interval [0, \( \infty \)]. Assuming that \( f \) is continuously differentiable, \( D \) is continuous. Furthermore, \( M \) is closed by Theorem 8.4.1. Therefore, the algorithmic map \( A \) is closed by Corollary 2 to Theorem 7.3.2. Finally, if \( x \in \Omega \), then \( \nabla f(x)^T d < 0 \), where \( d = -\nabla f(x) \). By Theorem 4.1.2, \( d \) is a descent direction, and hence \( f(y) < f(x) \) for \( y \in A(x) \). Assuming that the sequence generated by the algorithm is contained in a compact set, then by Theorem 7.2.3, the steepest descent algorithm converges to a point with zero gradient.

### Zigzagging of the Steepest Descent Method

The method of steepest descent usually works quite well during early stages of the optimization process, depending on the point of initialization. However, as a stationary point is approached, the method usually behaves poorly, taking small,
Figure 8.16 Method of steepest descent.

Table 8.11 Summary of Computations for the Method of Steepest Descent

<table>
<thead>
<tr>
<th>Iteration $k$</th>
<th>$\mathbf{x}_k$</th>
<th>$\nabla f(\mathbf{x}_k)$</th>
<th>$|\nabla f(\mathbf{x}_k)|$</th>
<th>$\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$</th>
<th>$\lambda_k$</th>
<th>$\mathbf{x}_{k+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.00, 3.00)</td>
<td>(−44.00, 24.00)</td>
<td>50.12</td>
<td>(44.00, −24.00)</td>
<td>0.062</td>
<td>(2.70, 1.51)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>52.00</td>
</tr>
<tr>
<td>2</td>
<td>(2.70, 1.51)</td>
<td>(0.73, 1.28)</td>
<td>1.47</td>
<td>(−0.73, −1.28)</td>
<td>0.24</td>
<td>(2.52, 1.20)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.34</td>
</tr>
<tr>
<td>3</td>
<td>(2.52, 1.20)</td>
<td>(0.80, −0.48)</td>
<td>0.93</td>
<td>(−0.80, 0.48)</td>
<td>0.11</td>
<td>(2.43, 1.25)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.09</td>
</tr>
<tr>
<td>4</td>
<td>(2.43, 1.25)</td>
<td>(0.18, 0.28)</td>
<td>0.33</td>
<td>(−0.18, −0.28)</td>
<td>0.31</td>
<td>(2.37, 1.16)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>5</td>
<td>(2.37, 1.16)</td>
<td>(0.30, −0.20)</td>
<td>0.36</td>
<td>(−0.30, 0.20)</td>
<td>0.12</td>
<td>(2.33, 1.18)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>(2.33, 1.18)</td>
<td>(0.08, 0.12)</td>
<td>0.14</td>
<td>(−0.08, −0.12)</td>
<td>0.36</td>
<td>(2.30, 1.14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>(2.30, 1.14)</td>
<td>(0.15, −0.08)</td>
<td>0.17</td>
<td>(−0.15, 0.08)</td>
<td>0.13</td>
<td>(2.28, 1.15)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.009</td>
</tr>
<tr>
<td>8</td>
<td>(2.28, 1.15)</td>
<td>(0.05, 0.08)</td>
<td>0.09</td>
<td></td>
<td>0.007</td>
<td></td>
</tr>
</tbody>
</table>
nearly orthogonal steps. This zigzagging phenomenon was encountered in Example 8.6.2 and is illustrated in Figure 8.16, in which zigzagging occurs along the valley shown by the dashed lines.

Zigzagging and poor convergence of the steepest descent algorithm at later stages can be explained intuitively by considering the following expression of the function $f$:

$$f(x_k + \lambda d) = f(x_k) + \lambda \nabla f(x_k)^T d + \lambda \|d\| \alpha(x_k; \lambda d),$$

where $\alpha(x_k; \lambda d) \to 0$ as $\lambda d \to 0$, and $d$ is a search direction with $\|d\| = 1$. If $x_k$ is close to a stationary point with zero gradient and $f$ is continuously differentiable, then $\|\nabla f(x_k)\|$ will be small, making the coefficient of $\lambda$ in the term $\lambda \nabla f(x_k)^T d$ of a small order of magnitude. Since the steepest descent method employs the linear approximation of $f$ to find a direction of movement, where the term $\lambda \|d\| \alpha(x_k; \lambda d)$ is essentially ignored, we should expect that the directions generated at late stages will not be very effective if the latter term contributes significantly to the description of $f$, even for relatively small values of $\lambda$.

As we shall learn in the remainder of the chapter, there are some ways to overcome the difficulties of zigzagging by deflecting the gradient. Rather than moving along $d = -\nabla f(x)$, we can move along $d = -D \nabla f(x)$ or along $d = -\nabla f(x) + g$, where $D$ is an appropriate matrix and $g$ is an appropriate vector. These correction procedures will be discussed in more detail shortly.

**Convergence Rate Analysis for the Steepest Descent Algorithm**

In this section we give a more formalized analysis of the zigzagging phenomenon and the empirically observed slow convergence rate of the steepest descent algorithm. This analysis will also afford insights into possible ways of alleviating this poor algorithmic performance.

Toward this end, let us begin by considering a bivariate quadratic function $f(x_1, x_2) = (1/2)(x_1^2 + \alpha x_2^2)$, where $\alpha > 1$. Note that the Hessian matrix to this function is $H = \text{diag}\{1, \alpha\}$, with eigenvalues 1 and $\alpha$. Let us define the condition number of a positive definite matrix to be the ratio of its largest to smallest eigenvalues. Hence, the condition number of $H$ for our example is $\alpha$. The contours of $f$ are plotted in Figure 8.17. Observe that as $\alpha$ increases, a phenomenon that is known as ill-conditioning, or a worsening of the condition number results, whereby the contours become increasingly skewed and the graph of the function becomes increasingly steep in the $x_2$ direction relative to the $x_1$ direction.
Chapter 8

Figure 8.17 Convergence rate analysis of the steepest descent algorithm.

Now, given a starting point \( x = (x_1, x_2)' \), let us apply an iteration of the steepest descent algorithm to obtain a point \( x_{\text{new}} = (x_{1\text{new}}, x_{2\text{new}})' \). Note that if \( x_1 = 0 \) or \( x_2 = 0 \), the procedure converges to the optimal minimizing solution \( x^* = (0, 0)' \) in one step. Hence, suppose that \( x_1 \neq 0 \) and \( x_2 \neq 0 \). The steepest descent direction is given by \( d = -\nabla f(x) = -(x_1, \alpha x_2)' \), resulting in \( x_{\text{new}} = x + \lambda d \), where \( \lambda \) solves the line search problem to minimize \( \theta(\lambda) = f(x + \lambda d) = (1/2)[x_1^2(1 - \lambda)^2 + \alpha x_2^2(1 - \alpha \lambda)^2] \) subject to \( \lambda \geq 0 \). Using simple calculus, we obtain

\[
\lambda = \frac{x_1^2 + \alpha^2 x_2^2}{x_1^2 + \alpha^3 x_2^2},
\]

so

\[
x_{\text{new}} = \begin{bmatrix}
\alpha^2 x_1 x_2^2 (\alpha - 1) \\
\alpha x_1^2 x_2^2 (1 - \alpha) \\
x_1^2 + \alpha^2 x_2^2 \\
x_1^2 + \alpha^3 x_2^2
\end{bmatrix} \begin{bmatrix}
x_1^2 \\
x_2^2 \\
x_1^2 + \alpha x_2^2 \\
x_1^2 + \alpha^2 x_2^2
\end{bmatrix}.
\]  

(8.13)

Observe that \( x_{1\text{new}}/x_{2\text{new}} = -\alpha^2 (x_2/x_1) \). Hence, if we begin with a solution \( x^0 \) having \( x_1^0/x_2^0 = K \neq 0 \) and generate a sequence of iterates \( \{x^k\}, k = 1, 2, \ldots, \) using the steepest descent algorithm, then the sequence of values \( \{x_1^k/x_2^k\} \) alternate between the values \( K \) and \( -\alpha^2/K \) as the sequence \( \{x^k\} \) converges to \( x^* = (0, 0)' \). For our example this means that the sequence zigzags between the pair of straight lines \( x_2 = (1/K)x_1 \) and \( x_2 = (-K/\alpha^2)x_1 \), as shown in Figure
8.17. Note that as the condition number $\alpha$ increases, this zigzagging phenomenon becomes more pronounced. On the other hand, if $\alpha = 1$, then the contours of $f$ are circular, and we obtain $x^1 = x^*$ in a single iteration.

To study the rate of convergence, let us examine the rate at which $\{f(x^k)\}$ converges to the value zero. From (8.13) it is easily verified that

$$
\frac{f(x^{k+1})}{f(x^k)} = \frac{K_k^2 \alpha (\alpha - 1)^2}{(K_k^2 + \alpha^3)(K_k^2 + \alpha)}, \quad \text{where } K_k = \frac{x_1^k}{x_2^k}. \quad (8.14)
$$

Indeed, the expression in (8.14) can be seen to be maximized when $K_k^2 = \alpha^2$ (see Exercise 8.19), so that we obtain

$$
\frac{f(x^{k+1})}{f(x^k)} \leq \frac{(\alpha - 1)^2}{(\alpha + 1)^2}. \quad (8.15)
$$

Note from (8.15) that $\{f(x^k)\} \to 0$ at a geometric or linear rate bounded by the ratio $(\alpha - 1)^2/(\alpha + 1)^2 < 1$. In fact, if we initialize the process with $x_1^0/x_2^0 = K = \alpha$, then, since $K_k^2 = (x_1^k/x_2^k)^2 = \alpha^2$ from above (see Figure 8.17), we get from (8.14) that the convergence ratio $f(x^{k+1})/f(x^k)$ is precisely $(\alpha - 1)^2/(\alpha + 1)^2$. Hence, as $\alpha$ approaches infinity, this ratio approaches 1 from below, and the rate of convergence becomes increasingly slower.

The foregoing analysis can be extended to a general quadratic function $f(x) = c^T x + (1/2)x^T H x$, where $H$ is an $n \times n$, symmetric, positive definite matrix. The unique minimizer $x^*$ for this function is given by the solution to the system $H x^* = -c$ obtained by setting $\nabla f(x^*) = 0$. Also, given an iterate $x_k$, the optimal step length $\lambda$ and the revised iterate $x_{k+1}$ are given by the following generalization of (8.13), where $g_k = \nabla f(x_k) = c + H x_k$:

$$
\lambda = \frac{g_k^T g_k}{g_k^T H g_k} \quad \text{and} \quad x_{k+1} = x_k - \lambda g_k. \quad (8.16)
$$

Now, to evaluate the rate of convergence, let us employ a convenient measure for convergence given by the following error function:

$$
e(x) = \frac{1}{2} (x - x^*)^T H (x - x^*) = f(x) + \frac{1}{2} x^T H x^*, \quad (8.17)
$$

where we have used the fact that $H x^* = -c$. Note that $e(x)$ differs from $f(x)$ by only a constant and equals zero if and only if $x = x^*$. In fact, it can be shown, analogous to (8.15), that (see Exercise 8.21)
where $\alpha$ is the condition number of $H$. Hence, $\{e(x_k)\} \to 0$ at a linear or geometric convergence rate bounded above by $(\alpha - 1)^2 / (\alpha + 1)^2$; so, as before, we can expect the convergence to become increasingly slower as $\alpha$ increases, depending on the initial solution $x_0$.

For continuously twice differentiable nonquadratic functions $f: \mathbb{R}^n \to \mathbb{R}$, a similar result is known to hold. In such a case, if $x^*$ is a local minimum to which a sequence $\{x_k\}$ generated by the steepest descent algorithm converges, and if $H(x^*)$ is positive definite with a condition number $\alpha$, then the corresponding sequence of objective values $\{f(x_k)\}$ can be shown to converge linearly to the value $f(x^*)$ at a rate bounded above by $(\alpha - 1)^2 / (\alpha + 1)^2$.

**Convergence Analysis of the Steepest Descent Algorithm Using Armijo’s Inexact Line Search**

In Section 8.3 we introduced Armijo’s rule for selecting an acceptable, inexact step length during a line search process. It is instructive to observe how such a criterion still guarantees algorithmic convergence. Below, we present a convergence analysis for an inexact steepest descent algorithm applied to a function $f: \mathbb{R}^n \to \mathbb{R}$ whose gradient function $\nabla f(x)$ is Lipschitz continuous with constant $G > 0$ on $S(x_0) = \{x : f(x) \leq f(x_0)\}$ for some given $x_0 \in \mathbb{R}^n$. That is, we have $\|\nabla f(x) - \nabla f(y)\| \leq G \|x - y\|$ for all $x, y \in S(x_0)$. For example, if the Hessian of $f$ at any point has a norm bounded above by a constant $G$ on $\text{conv} S(x_0)$ (see Appendix A for the norm of a matrix), then such a function has Lipschitz continuous gradients. This follows from the mean value theorem, noting that for any $x \neq y \in S(x_0)$, $\|\nabla f(x) - \nabla f(y)\| = \|H(\tilde{x})(x - y)\| \leq G \|x - y\|$.

The procedure we analyze is the often-used variant of Armijo’s rule described in Section 8.3 with parameters $0 < \varepsilon < 1$, $\alpha = 2$, and a fixed-step-length parameter $\lambda$, wherein either $\lambda$ itself is chosen, if acceptable, or is sequentially halved until an acceptable step length results. This procedure is embodied in the following result.

**8.6.3 Theorem**

Let $f: \mathbb{R}^n \to \mathbb{R}$ be such that its gradient $\nabla f(x)$ is Lipschitz continuous with constant $G > 0$ on $S(x_0) = \{x : f(x) \leq f(x_0)\}$ for some given $x_0 \in \mathbb{R}^n$. Pick some fixed-step-length parameter $\lambda > 0$, and let $0 < \varepsilon < 1$. Given any iterate
Unconstrained Optimization

393

\( x_k \), define the search direction \( d_k = -\nabla f(x_k) \), and consider Armijo’s function
\[ \hat{\theta}(\lambda) = \theta(0) + \lambda \theta'(0), \quad \lambda \geq 0, \]
where \( \theta(\lambda) = f(x_k + \lambda d_k) \), \( \lambda \geq 0 \), is the line search function. If \( d_k = 0 \), then stop. Otherwise, find the smallest integer \( t \geq 0 \) for which
\[ \theta(\lambda / 2^t) \leq \hat{\theta}(\lambda / 2^t) \]
and define the next iterate as
\[ x_{k+1} = x_k + \lambda_k d_k, \]
where \( \lambda_k = \lambda / 2^t \). Now suppose that starting with some iterate \( x_0 \), this procedure produces a sequence of iterates \( x_0, x_1, x_2, \ldots \). Then either the procedure terminates finitely with \( \nabla f(x_k) = 0 \) for some \( K \), or else an infinite sequence \( \{x_k\} \) is generated such that the corresponding sequence \( \{\nabla f(x_k)\} \to 0 \).

**Proof**

The case of finite termination is clear. Hence, suppose that an infinite sequence \( \{x_k\} \) is generated. Note that the Armijo criterion \( \theta(\lambda / 2^t) \leq \hat{\theta}(\lambda / 2^t) \) is equivalent to
\[ \theta(\lambda / 2^t) = f(x_{k+1}) \leq \hat{\theta}(\lambda / 2^t) = \theta(0) + (\lambda \theta'/2^t) \nabla f(x_k)'d_k = f(x_k) - (\lambda \theta'/2^t) \|\nabla f(x_k)\|^2. \]
Hence, \( t \geq 0 \) is the smallest integer for which
\[ f(x_{k+1}) - f(x_k) \leq -\frac{\lambda \theta}{2^t} \|\nabla f(x_k)\|^2. \]  \( (8.19) \)

Now, using the mean value theorem, we have, for some strict convex combination \( \bar{x} \) of \( x_k \) and \( x_{k+1} \), that
\[
\begin{align*}
    f(x_{k+1}) - f(x_k) &= \lambda_k d_k' \nabla f(\bar{x}) \\
    &= -\lambda_k \nabla f(x_k)'[\nabla f(x_k) - \nabla f(\bar{x})] \\
    &= -\lambda_k \|\nabla f(x_k)\|^2 + \lambda_k \nabla f(x_k)'[\nabla f(x_k) - \nabla f(\bar{x})] \\
    &\leq -\lambda_k \|\nabla f(x_k)\|^2 + \lambda_k \|\nabla f(x_k)\| \|\nabla f(x_k) - \nabla f(\bar{x})\|.
\end{align*}
\]
But by the Lipschitz continuity of \( \nabla f \), noting from \( (8.19) \) that the descent nature of the algorithm guarantees that \( x_k \in S(x_0) \) for all \( k \), we have
\[ \|\nabla f(x_k) - \nabla f(\bar{x})\| \leq G \|x_k - \bar{x}\| \leq G \|x_k - x_{k+1}\| = G \lambda_k \|\nabla f(x_k)\|. \]
Substituting this above, we obtain
\[ f(x_{k+1}) - f(x_k) \leq -\lambda_k \|\nabla f(x_k)\|^2(1 - \lambda_k G) = \frac{-\lambda}{2^t} \|\nabla f(x_k)\|^2 \left( 1 - \frac{\lambda G}{2^t} \right). \]  \( (8.20) \)
Consequently, from \( (8.20) \), we know that \( (8.19) \) will hold true when \( t \) is increased to no larger an integer value than is necessary to make \( 1 - (\lambda G/2^t) \geq \)
$\epsilon$, for then (8.20) will imply (8.19). But this means that $1 - (\lambda G / 2^{t-1}) < \epsilon$; that is, $\lambda \epsilon / 2^t > (1 - \epsilon) / 2G$. Substituting this in (8.19), we get

$$f(x_{k+1}) - f(x_k) < -\epsilon (1 - \epsilon) \| \nabla f(x_k) \|^2 / 2G.$$ 

Hence, noting that $\{f(x_k)\}$ is a monotone decreasing sequence and so has a limit, taking limits as $t \to \infty$, we get

$$0 \leq -\epsilon (1 - \epsilon) \lim_{k \to \infty} \| \nabla f(x_k) \|^2 / 2G,$$

which implies that $\{\nabla f(x_k)\} \to 0$. This completes the proof.

**Method of Newton**

In Section 8.2 we discussed Newton's method for minimizing a function of a single variable. The method of Newton is a procedure that deflects the steepest descent direction by premultiplying it by the inverse of the Hessian matrix. This operation is motivated by finding a suitable direction for the quadratic approximation to the function rather than by finding a linear approximation to the function, as in the gradient search. To motivate the procedure, consider the following approximation $q$ at a given point $x_k$:

$$q(x) = f(x_k) + \nabla f(x_k)'(x - x_k) + \frac{1}{2}(x - x_k)'H(x_k)(x - x_k),$$

where $H(x_k)$ is the Hessian matrix of $f$ at $x_k$. A necessary condition for a minimum of the quadratic approximation $q$ is that $\nabla q(x) = 0$, or $\nabla f(x_k)' + H(x_k)(x - x_k) = 0$. Assuming that the inverse of $H(x_k)$ exists, the successor point $x_{k+1}$ is given by

$$x_{k+1} = x_k - H(x_k)^{-1}\nabla f(x_k). \quad (8.21)$$

Equation (8.21) gives the recursive form of the points generated by Newton's method for the multidimensional case. Assuming that $\nabla f(\bar{x}) = 0$, that $H(\bar{x})$ is positive definite at a local minimum $\bar{x}$, and that $f$ is continuously twice differentiable, it follows that $H(x_k)$ is positive definite at points close to $\bar{x}$, and hence the successor point $x_{k+1}$ is well defined.

It is interesting to note that Newton's method can be interpreted as a steepest descent algorithm with affine scaling. Specifically, given a point $x_k$ at iteration $k$, suppose that $H(x_k)$ is positive definite and that we have a Cholesky factorization (see Appendix A.2) of its inverse given by $H(x_k)^{-1} = LL'$, where $L$ is a lower triangular matrix with positive diagonal elements. Now, consider
the affine scaling transformation $x = Ly$. This transforms the function $f(x)$ to the function $F(y) = f(Ly)$, and the current point in the $y$ space is $y_k = L^{-1}x_k$.

Hence, we have $\nabla F(y_k) = L'\nabla f(Ly_k) = L'\nabla f(x_k)$. A unit step size along the negative gradient direction in the $y$ space will then take us to the point $y_{k+1} = y_k - L'\nabla f(x_k)$. Translating this to the corresponding movement in the $x$ space by premultiplying throughout by $L$ produces precisely Equation (8.21) and hence yields a steepest descent interpretation of Newton’s method. Observe that this comment alludes to the benefits of using an appropriate scaling transformation. Indeed, if the function $f$ was quadratic in the above analysis, then a unit step along the steepest descent direction in the transformed space would be an optimal step along that direction, which would moreover take us directly to the optimal solution in one iteration starting from any given solution.

We also comment here that (8.21) can be viewed as an application of the Newton–Raphson method to the solution of the system of equations $\nabla f(x) = 0$.

Given a well-determined system of nonlinear equations, each iteration of the Newton–Raphson method adopts a first-order Taylor series approximation to this equation system at the current iterate and solves the resulting linear system to determine the next iterate. Applying this to the system $\nabla f(x) = 0$ at an iterate $x_k$, the first-order approximation to $\nabla f(x)$ is given by $\nabla f(x_k) + H(x_k)(x - x_k)$. Setting this equal to zero and solving produces the solution $x = x_{k+1}$ as given by (8.21).

### 8.6.4 Example

Consider the following problem:

$$
\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.
$$

The summary of the computations using Newton’s method is given in Table 8.12. At each iteration, $x_{k+1}$ is given by $x_{k+1} = x_k - H(x_k)^{-1}\nabla f(x_k)$. After six iterations, the point $x_7 = (1.83, 0.91)^T$ is reached. At this point, $\|\nabla f(x_7)\| = 0.04$, and the procedure is terminated. The points generated by the method are shown in Figure 8.18.

In Example 8.6.4 the value of the objective function decreased at each iteration. However, this will not generally be the case, so $f$ cannot be used as a descent function. Theorem 8.6.5 indicates that Newton’s method indeed converges, provided that we start from a point close enough to an optimal point.

### Order-Two Convergence of the Method of Newton

In general, the points generated by the method of Newton may not converge. The reason for this is that $H(x_k)$ may be singular, so that $x_{k+1}$ is not
Table 8.12 Summary of Computations for the Method of Newton

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$x_k$</th>
<th>$f(x_k)$</th>
<th>$\nabla f(x_k)$</th>
<th>$H(x_k)$</th>
<th>$H(x_k)^{-1}$</th>
<th>$-H(x_k)^{-1}\nabla f(x_k)$</th>
<th>$x_{k+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.00, 3.00)</td>
<td>52.00</td>
<td>(-44.0, 24.0)</td>
<td>$\begin{bmatrix} 50.0 &amp; -4.0 \ -4.0 &amp; 8.0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 8.0 &amp; 4.0 \ 4.0 &amp; 50.0 \end{bmatrix}$</td>
<td>(0.67, -2.67)</td>
<td>(0.67, 0.33)</td>
</tr>
<tr>
<td>2</td>
<td>(0.67, 0.33)</td>
<td>3.13</td>
<td>(-9.39, -0.04)</td>
<td>$\begin{bmatrix} 23.23 &amp; -4.0 \ -4.0 &amp; 8.0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 8.0 &amp; 4.0 \ 4.0 &amp; 23.23 \end{bmatrix}$</td>
<td>(0.44, 0.23)</td>
<td>(1.11, 0.56)</td>
</tr>
<tr>
<td>3</td>
<td>(1.11, 0.56)</td>
<td>0.63</td>
<td>(-2.84, -0.04)</td>
<td>$\begin{bmatrix} 11.50 &amp; -4.0 \ -4.0 &amp; 8.0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 8.0 &amp; 4.0 \ 4.0 &amp; 11.50 \end{bmatrix}$</td>
<td>(0.30, 0.14)</td>
<td>(1.41, 0.70)</td>
</tr>
<tr>
<td>4</td>
<td>(1.41, 0.70)</td>
<td>0.12</td>
<td>(-0.80, -0.04)</td>
<td>$\begin{bmatrix} 6.18 &amp; -4.0 \ -4.0 &amp; 8.0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 8.0 &amp; 4.0 \ 4.0 &amp; 6.18 \end{bmatrix}$</td>
<td>(0.20, 0.10)</td>
<td>(1.61, 0.80)</td>
</tr>
<tr>
<td>5</td>
<td>(1.61, 0.80)</td>
<td>0.02</td>
<td>(-0.22, -0.04)</td>
<td>$\begin{bmatrix} 3.83 &amp; -4.0 \ -4.0 &amp; 8.0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 8.0 &amp; 4.0 \ 4.0 &amp; 3.83 \end{bmatrix}$</td>
<td>(0.13, 0.07)</td>
<td>(1.74, 0.87)</td>
</tr>
<tr>
<td>6</td>
<td>(1.74, 0.87)</td>
<td>0.005</td>
<td>(-0.07, 0.00)</td>
<td>$\begin{bmatrix} 2.81 &amp; -4.0 \ -4.0 &amp; 8.0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 8.0 &amp; 4.0 \ 4.0 &amp; 2.81 \end{bmatrix}$</td>
<td>(0.09, 0.04)</td>
<td>(1.83, 0.91)</td>
</tr>
<tr>
<td>7</td>
<td>(1.83, 0.91)</td>
<td>0.0009</td>
<td>(-0.003, -0.04)</td>
<td>$\begin{bmatrix} 2.81 &amp; -4.0 \ -4.0 &amp; 8.0 \end{bmatrix}$</td>
<td>$\begin{bmatrix} 8.0 &amp; 4.0 \ 4.0 &amp; 2.81 \end{bmatrix}$</td>
<td>(0.09, 0.04)</td>
<td>(1.83, 0.91)</td>
</tr>
</tbody>
</table>
well defined. Even if $H(x_k)^{-1}$ exists, $f(x_{k+1})$ is not necessarily less than $f(x_k)$. However, if the starting point is close enough to a point $\bar{x}$ such that $\nabla f(\bar{x}) = 0$ and $H(\bar{x})$ is of full rank, then the method of Newton is well defined and converges to $\bar{x}$. This is proved in Theorem 8.6.5 by showing that all the assumptions of Theorem 7.2.3 hold true, where the descent function $\alpha$ is given by $\alpha(x) = \|x - \bar{x}\|$.

**8.6.5 Theorem**

Let $f: \mathbb{R}^n \to \mathbb{R}$ be continuously twice differentiable. Consider Newton's algorithm defined by the map $A(x) = x - H(x)^{-1}\nabla f(x)$. Let $\bar{x}$ be such that $\nabla f(\bar{x}) = 0$ and $H(\bar{x})^{-1}$ exists. Let the starting point $x_1$ be sufficiently close to $\bar{x}$ so that this proximity implies that there exist $k_1, k_2 > 0$ with $k_1k_2\|x_1 - \bar{x}\| < 1$ such that
1. \[ \|H(\bar{x})\|^{\dagger} \leq k_1 \]

and by the Taylor series expansion of \( \nabla f \),

2. \[ \|\nabla f(\bar{x}) - \nabla f(x) - H(x)(\bar{x} - x)\| \leq k_2 \|\bar{x} - x\|^2 \]

for each \( x \) satisfying \( \|x - \bar{x}\| \leq \|x_1 - \bar{x}\| \). Then the algorithm converges superlinearly to \( \bar{x} \) with at least an order-two or quadratic rate of convergence.

**Proof**

Let the solution set \( \Omega = \{\bar{x}\} \) and let \( X = \{x : \|x - \bar{x}\| \leq \|x_1 - \bar{x}\|\} \). We prove convergence by using Theorem 7.2.3. Note that \( X \) is compact and that the map \( A \) given via (8.21) is closed on \( X \). We now show that \( \alpha(x) = \|x - \bar{x}\| \) is indeed a descent function. Let \( x \in X \), and suppose that \( x \neq \bar{x} \). Let \( y \in A(x) \). Then, by the definition of \( A \) and since \( \nabla f(\bar{x}) = 0 \), we get

\[
y - \bar{x} = (x - \bar{x}) - H(x)^{-1}[\nabla f(x) - \nabla f(\bar{x})] = H(x)^{-1}[\nabla f(\bar{x}) - \nabla f(x) - H(x)(\bar{x} - x)].
\]

Noting 1 and 2, it then follows that

\[
\|y - \bar{x}\| = \|H(x)^{-1}[\nabla f(\bar{x}) - \nabla f(x) - H(x)(\bar{x} - x)]\|
\]

\[
\leq \|H(x)^{-1}\|[\nabla f(\bar{x}) - \nabla f(x) - H(x)(\bar{x} - x)]
\]

\[
\leq k_1 k_2 \|x - \bar{x}\|^2 \leq k_1 k_2 \|x_1 - \bar{x}\|\|x - \bar{x}\|
\]

\[
< \|x - \bar{x}\|.
\]

This shows that \( \alpha \) is indeed a descent function. By the corollary to Theorem 7.2.3, we have convergence to \( \bar{x} \). Moreover, for any iterate \( x_k \in X \), the new iterate \( y = x_{k+1} \) produced by the algorithm satisfies \( \|x_{k+1} - \bar{x}\| \leq k_1 k_2 \|x_k - \bar{x}\|^2 \) from above. Since \( \{x_k\} \to \bar{x} \), we have at least an order-two rate of convergence.

### 8.7 Modification of Newton’s Method: Levenberg–Marquardt and Trust Region Methods

In Theorem 8.6.5 we have seen that if Newton's method is initialized close enough to a local minimum \( \bar{x} \) with a positive definite Hessian \( H(\bar{x}) \), then it converges quadratically to this solution. In general, we have observed that the

---

\( \dagger \) See Appendix A.1 for the norm of a matrix.
method may not be defined because of the singularity of \( H(x_k) \) at a given point \( x_k \), or the search direction \( d_k = -H(x_k)^{-1}Vf(x_k) \) may not be a descent direction; or even if \( \nabla f(x_k)^T d_k < 0 \), a unit step size might not give a descent in \( f \). To safeguard against the latter, we could perform a line search given that \( d_k \) is a descent direction. However, for the more critical issue of having a well-defined algorithm that converges to a point of zero gradient irrespective of the starting solution (i.e., enjoys *global convergence*), the following modifications can be adopted.

We first discuss a modification of Newton’s method that guarantees convergence regardless of the starting point. Given \( x \), consider the direction \( d = -BVf(x) \), where \( B \) is a symmetric positive definite matrix to be determined later. The successor point is \( y = x + \lambda d \), where \( \lambda \) is an optimal solution to the problem to minimize \( f(x + \lambda d) \) subject to \( \lambda \geq 0 \).

We now specify the matrix \( B \) as \( (\epsilon I + H)^{-1} \), where \( H = H(x) \). The scalar \( \epsilon \geq 0 \) is determined as follows. Fix \( \delta > 0 \), and let \( \epsilon \geq 0 \) be the smallest scalar that would make all the eigenvalues of the matrix \( (\epsilon I + H) \) greater than or equal to \( \delta \). Since the eigenvalues of \( \epsilon I + H \) are all positive, \( \epsilon I + H \) is positive definite and invertible. In particular, \( B = (\epsilon I + H)^{-1} \) is also positive definite. Since the eigenvalues of a matrix depend continuously on its elements, \( \epsilon \) is a continuous function of \( x \), and hence the point-to-point map \( D: \mathbb{R}^n \to \mathbb{R}^n \) defined by \( D(x) = (x, d) \) is continuous. Thus, the algorithmic map is \( A = MD \), where \( M \) is the usual line search map over \( \{A: A \geq 0\} \).

Let \( \Omega = \{\bar{x}: \nabla f(\bar{x}) = 0\} \), and let \( x \notin \Omega \). Since \( B \) is positive definite, \( d = -BVf(x) \neq 0 \); and, by Theorem 8.4.1, it follows that \( M \) is closed at \( (x, d) \). Furthermore, since \( D \) is a continuous function, by Corollary 2 to Theorem 7.3.2, \( A = MD \) is closed over the complement of \( \Omega \).

To invoke Theorem 7.2.3, we need to specify a continuous descent function. Suppose that \( x \notin \Omega \), and let \( y \in A(x) \). Note that \( \nabla f(x)^T d = -\nabla f(x)^T BVf(x) < 0 \) since \( B \) is positive definite and \( \nabla f(x) \neq 0 \). Thus, \( d \) is a descent direction of \( f \) at \( x \), and by Theorem 4.1.2, \( f(y) < f(x) \). Therefore, \( f \) is indeed a descent function. Assuming that the sequence generated by the algorithm is contained in a compact set, by Theorem 7.2.3 it follows that the algorithm converges.

It should be noted that if the smallest eigenvalue of \( H(\bar{x}) \) is greater than or equal to \( \delta \), then, as the points \( \{x_k\} \) generated by the algorithm approach \( \bar{x} \), \( \epsilon_k \) will be equal to zero. Thus, \( d_k = -H(x_k)^{-1}Vf(x_k) \), and the algorithm reduces to that of Newton and, hence, this method also enjoys an order-two rate of convergence.
This underscores the importance of selecting $\delta$ properly. If $\delta$ is chosen to be too small to ensure the asymptotic quadratic convergence rate because of the reduction of the method to Newton’s algorithm, ill-conditioning might occur at points where the Hessian is (near) singular. On the other hand, if $\delta$ is chosen to be very large, which would necessitate using a large value of $\epsilon$ and would make $B$ diagonally dominant, the method would behave similar to the steepest descent algorithm, and only a linear convergence rate would be realized.

The foregoing algorithmic scheme of determining the new iterate $x_{k+1}$ from an iterate $x_k$ according to the solution of the system

\[ [\delta I + H(x_k)](x_{k+1} - x_k) = -\nabla f(x_k) \]  

(8.22)

in lieu of (8.21) is generally known as a Levenberg–Marquardt method, following a similar scheme proposed for solving nonlinear least squares problems. A typical operational prescription for such a method is as follows. (The parameters 0.25, 0.75, 2, 4, etc., used below have been found to work well empirically, and the method is relatively insensitive to these parameter values.)

Given an iterate $x_k$ and a parameter $\epsilon_k > 0$, first ascertain the positive definiteness of $\epsilon_k I + H(x_k)$ by attempting to construct its Cholesky factorization $LL'$ (see Appendix A.2). If this is unsuccessful, then multiply $\epsilon_k$ by a factor of 4 and repeat until such a factorization is available. Then solve the system (8.22) via $LL'(x_{k+1} - x_k) = -\nabla f(x_k)$, exploiting the triangularity of $L$ to obtain $x_{k+1}$. Compute $f(x_{k+1})$ and determine $R_k$ as the ratio of the actual decrease $f(x_k) - f(x_{k+1})$ in $f$ to its predicted decrease $q(x_k) - q(x_{k+1})$ as foretold by the quadratic approximation $q$ to $f$ at $x = x_k$. Note that the closer $R_k$ is to unity, the more reliable is the quadratic approximation, and the smaller we can afford $\epsilon$ to be. With this motivation, if $R_k < 0.25$, put $\epsilon_{k+1} = 4\epsilon_k$; if $R_k > 0.75$, put $\epsilon_{k+1} = \epsilon_k/2$; otherwise, put $\epsilon_{k+1} = \epsilon_k$. Furthermore, in case $R_k \leq 0$ so that no improvement in $f$ is realized, reset $x_{k+1} = x_k$; or else, retain the computed $x_{k+1}$. Increment $k$ by 1 and reiterate until convergence to a point of zero gradient is obtained.

A scheme of this type bears a close resemblance and relationship to trust region methods, or restricted step methods, for minimizing $f$. Note that the main difficulty with Newton’s method is that the region of trust within which the quadratic approximation at a given point $x_k$ can be considered to be sufficiently reliable might not include a point in the solution set. To circumvent this problem, we can consider the trust region subproblem:

\[ \text{Minimize } \{q(x): x \in \Omega_k\}, \]  

(8.23)

where $q$ is the quadratic approximation to $f$ at $x = x_k$ and $\Omega_k$ is a trust region defined by $\Omega_k = \{x: \|x - x_k\| \leq \Delta_k\}$ for some trust region parameter $\Delta_k > 0$. 

(Here $\| \cdot \|$ is the $\ell_2$ norm; when the $\ell_\infty$ norm is used instead, the method is also known as the box-step, or hypercube, method.) Now let $x_{k+1}$ solve (8.23) and, as before, define $R_k$ as the ratio of the actual to the predicted descent. If $R_k$ is too small relative to unity, then the trust region needs to be reduced; but if it is sufficiently respectable in value, the trust region can actually be expanded. The following is a typical prescription for defining $\Delta_{k+1}$ for the next iteration, where again, the method is known to be relatively insensitive to the specified parameter choices. If $R_k < 0.25$, put $\Delta_{k+1} = \|x_{k+1} - x_k\|/4$. If $R_k > 0.75$ and $\|x_{k+1} - x_k\| = \Delta_k$, that is, the trust region constraint is binding in (8.23), then put $\Delta_{k+1} = 2\Delta_k$. Otherwise, retain $\Delta_{k+1} = \Delta_k$. Furthermore, in case $R_k \leq 0$ so that $f$ did not improve at this iteration, reset $x_{k+1}$ to $x_k$ itself. Then increment $k$ by 1 and repeat until a point with a zero gradient obtains. If this does not occur finitely, it can be shown that if the sequence $\{x_k\}$ generated is contained in a compact set, and if $f$ is continuously twice differentiable, then there exists an accumulation point $\bar{x}$ of this sequence for which $\nabla f(\bar{x}) = 0$ and $H(\bar{x})$ is positive semidefinite. Moreover, if $H(\bar{x})$ is positive definite, then for $k$ sufficiently large, the trust region bound is inactive, and hence, the method reduces to Newton’s method with a second-order rate of convergence (see the Notes and References section for further details).

There are two noteworthy points in relation to the foregoing discussion. First, wherever the actual Hessian has been employed above in the quadratic representation of $f$, an approximation to this Hessian can be used in practice, following quasi-Newton methods as discussed in the next section. Second, observe that by writing $\delta = x - x_k$ and equivalently, squaring both sides of the constraint defining $R_k$, we can write (8.23) explicitly as follows:

$$\text{Minimize } \left\{ \nabla f(x_k)' \delta + \frac{1}{2} \delta' H(x_k) \delta : \frac{1}{2} \| \delta \|^2 \leq \frac{1}{2} \Delta_k^2 \right\}. \quad (8.24)$$

The KKT conditions for (8.24) require a nonnegative Lagrange multiplier $\lambda$ and a primal feasible solution $\delta$ such that the following holds true in addition to the complementary slackness condition:

$$[H(x_k) + \lambda I] \delta = -\nabla f(x_k).$$

Note the resemblance of this to the Levenberg–Marquardt method given by (8.22). In particular, if $\Delta_k = -[H(x_k) + \varepsilon_k I]^{-1} \nabla f(x_k)$ in (8.24), where $H(x_k) + \varepsilon_k I$ is positive definite, then, indeed, it is readily verified that $\delta = x_{k+1} - x_k$ given by (8.22) and $\lambda = \varepsilon_k$ satisfy the saddle point optimality conditions for (8.24) (see Exercise 8.29). Hence, the Levenberg–Marquardt scheme described above can be viewed as a trust region type of method as well.
Finally, let us comment on a *dog-leg trajectory* proposed by Powell, which more directly follows the philosophy described above of compromising between a steepest descent step and Newton’s step, depending on the trust region size $\Delta_k$. Referring to Figure 8.19, let $x_{k+1}^{\text{SD}}$ and $x_{k+1}^{\text{N}}$, respectively, denote the new iterate obtained via a steepest descent step, (8.16), and a Newton step, (8.21) ($x_{k+1}^{\text{SD}}$ is sometimes also called the *Cauchy point*). The piecewise linear curve defined by the line segments joining $x_k$ to $x_{k+1}^{\text{SD}}$ and $x_{k+1}^{\text{SD}}$ to $x_{k+1}^{\text{N}}$ is called the *dog-leg trajectory*. It can be shown that along this trajectory, the distance from $x_k$ increases monotonically while the objective value of the quadratic model falls. The proposed new iterate $x_{k+1}^{\text{N}}$ is taken as the (unique) point at which the circle with radius $\Delta_k$ and centered at $x_k$ intercepts this trajectory, if at all, as shown in Figure 8.19, and is taken as the Newton iterate $x_{k+1}^{\text{N}}$ otherwise. Hence, when $\Delta_k$ is small relative to the dog-leg trajectory, the method behaves as a steepest descent algorithm; and with a relatively larger $\Delta_k$, it reduces to Newton’s method. Again, under suitable assumptions, as above, second-order convergence to a stationary point can be established. Moreover, the algorithmic step is simple and obviates (8.22) or (8.23). We refer the reader to the Notes and References section for further reading on this subject.

### 8.8 Methods Using Conjugate Directions: Quasi-Newton and Conjugate Gradient Methods

In this section we discuss several procedures that are based on the important concept of conjugacy. Some of these procedures use derivatives, whereas others use only functional evaluations. The notion of conjugacy defined below is very useful in unconstrained optimization. In particular, if the objective function is quadratic, then, by searching along conjugate directions, in any order, the minimum point can be obtained in, at most, $n$ steps.

![Figure 8.19 Dog-leg trajectory.](image)
8.8.1 Definition

Let $H$ be an $n \times n$ symmetric matrix. The vectors $d_1, \ldots, d_n$ are called $H$-conjugate or simply conjugate if they are linearly independent and if $d_i^T H d_j = 0$ for $i \neq j$.

It is instructive to observe the significance of conjugacy to the minimization of quadratic functions. Consider the quadratic function $f(x) = c^T x + (1/2) x^T H x$, where $H$ is an $n \times n$ symmetric matrix, and suppose that $d_1, \ldots, d_n$ are $H$-conjugate directions. By the linear independence of these direction vectors, given a starting point $x_1$, any point $x$ can be uniquely represented as $x = x_1 + \sum_{j=1}^n \lambda_j d_j$. Using this substitution we can rewrite $f(x)$ as the following function of $\lambda$:

$$
f(x) = c^T x_1 + \sum_{j=1}^n \lambda_j c^T d_j + \frac{1}{2} \left( x_1 + \sum_{j=1}^n \lambda_j d_j \right)^T H \left( x_1 + \sum_{j=1}^n \lambda_j d_j \right).
$$

Using the $H$-conjugacy of $d_1, \ldots, d_n$, this simplifies equivalently to minimizing

$$
F(\lambda) = \sum_{j=1}^n \left[ c^T (x_1 + \lambda_j d_j) + \frac{1}{2} (x_1 + \lambda_j d_j)^T H (x_1 + \lambda_j d_j) \right].
$$

Observe that $F$ is separable in $\lambda_1, \ldots, \lambda_n$ and can be minimized by minimizing each term in $[*]$ independently and then composing the net result. Note that the minimization of each such term corresponds to minimizing $f$ from $x_1$ along the direction $d_j$. (In particular, if $H$ is positive definite, the minimizing value of $\lambda_j$ is given by $\lambda_j^* = -[c^T d_j + x_1^T H d_j] / d_j^T H d_j$ for $j = 1, \ldots, n$. Alternatively, the foregoing derivation readily reveals that the same minimizing step lengths $\lambda_j^*$, $j = 1, \ldots, n$, result if we sequentially minimize $f$ from $x_1$ along the directions $d_1, \ldots, d_n$ in any order, leading to an optimal solution.

The following example illustrates the notion of conjugacy and highlights the foregoing significance of optimizing along conjugate directions for quadratic functions.

8.8.2 Example

Consider the following problem:

Minimize $-12x_2 + 4x_1^2 + 4x_2^2 + 4x_1 x_2$.

Note that the Hessian matrix $H$ is given by
We now generate two conjugate directions, \( d_1 \) and \( d_2 \). Suppose that we choose \( d_1^f = (1, 0) \). Then \( d_2^f = (a, b) \) must satisfy \( 0 = \langle d_1^f H d_2, d_2^f \rangle = 8a - 4b \). In particular, we may choose \( a = 1 \) and \( b = 2 \) so that \( d_2^f = (1, 2) \). It may be noted that the conjugate directions are not unique.

If we minimize the objective function \( f \) starting from \( x_1^f = (-1/2, 1) \) along the direction \( d_1 \), we get the point \( x_2^f = (1/2, 1) \). Now, starting from \( x_2 \) and minimizing along \( d_2 \), we get \( x_3^f = (1, 2) \). Note that \( x_3 \) is the minimizing point.

The contours of the objective function and the path taken to reach the optimal point are shown in Figure 8.20. The reader can easily verify that starting from any point and minimizing along \( d_1 \) and \( d_2 \), the optimal point is reached in, at most, two steps. For example, the dashed lines in Figure 8.20 exhibit the path obtained by sequentially minimizing along another pair of conjugate directions. Furthermore, if we had started at \( x_1 \) and then minimized along \( d_2 \) first and next along \( d_1 \), the optimizing step lengths along these respective directions would have remained the same as for the first case, taking the iterates from \( x_1 \) to \( x_2^f = (0, 2)^f \) to \( x_3 \).
Optimization of Quadratic Functions: Finite Convergence

Example 8.8.2 demonstrates that a quadratic function can be minimized in, at most, \( n \) steps whenever we search along conjugate directions of the Hessian matrix. This result is generally true for quadratic functions, as shown by Theorem 8.8.3. This, coupled with the fact that a general function can be closely represented by its quadratic approximation in the vicinity of the optimal point, makes the notion of conjugacy very useful for optimizing both quadratic and nonquadratic functions. Note also that this result shows that if we start at \( x_1 \), then at each step \( k = 1, \ldots, n \), the point \( x_{k+1} \) obtained minimizes \( f \) over the linear subspace containing \( x_1 \) that is spanned by the vectors \( d_1, \ldots, d_k \). Moreover, the gradient \( \nabla f(x_{k+1}) \), if nonzero, is orthogonal to this subspace. This is sometimes called the expanding subspace property and is illustrated in Figure 8.21 for \( k = 1, 2 \).

**8.8.3 Theorem**

Let \( f(x) = c^t x + (1/2) x^t H x \), where \( H \) is an \( n \times n \) symmetric matrix. Let \( d_1, \ldots, d_n \) be \( H \)-conjugate, and let \( x_1 \) be an arbitrary starting point. For \( k = 1, \ldots, n \), let \( \lambda_k \) be an optimal solution to the problem to minimize \( f(x_k + \lambda d_k) \) subject to \( \lambda \in \mathbb{R} \), and let \( x_{k+1} = x_k + \lambda_k d_k \). Then, for \( k = 1, \ldots, n \), we must have:

1. \( \nabla f(x_{k+1})^t d_j = 0 \) for \( j = 1, \ldots, k \).
2. \( \nabla f(x_1)^t d_k = \nabla f(x_k)^t d_k \).
3. \( x_{k+1} \) is an optimal solution to the problem to minimize \( f(x) \) subject to \( x - x_1 \in L(d_1, \ldots, d_k) \), where \( L(d_1, \ldots, d_k) \) is the linear subspace formed by \( d_1, \ldots, d_k \); that is, \( L(d_1, \ldots, d_k) = \{ \sum_{j=1}^k \mu_j d_j : \mu_j \in \mathbb{R} \text{ for each } j \} \). In particular, \( x_{n+1} \) is a minimizing point of \( f \) over \( \mathbb{R}^n \).

---

![Figure 8.21 Expanding subspace property.](image)
To prove Part 1, first note that $f(x_j + \lambda d_j)$ achieves a minimum at $\lambda_j$ only if $\nabla f(x_j + \lambda d_j)'d_j = 0$; that is, $\nabla f(x_{j+1})'d_j = 0$. Thus, Part 1 holds true for $j = k$. For $j < k$, note that

$\nabla f(x_{k+1}) = c + Hx_{j+1} = c + Hx_{j+1} + H \left( \sum_{i=j+1}^{k} \lambda_i d_i \right) = \nabla f(x_{j+1}) + H \left( \sum_{i=j+1}^{k} \lambda_i d_i \right)$.  \hspace{1cm} (8.25)

By conjugacy, $d_i'Hd_j = 0$ for $i = j + 1, \ldots, k$. Thus, from (8.25) it follows that $\nabla f(x_{k+1})'d_j = 0$, and Part 1 holds true.

Replacing $k$ by $k - 1$ and letting $j = 0$ in (8.25), we get

\[
\nabla f(x_k) = \nabla f(x_1) + H \left( \sum_{i=1}^{k-1} \lambda_i d_i \right) \quad \text{for } k \geq 2.
\]

Multiplying by $d_i'$ and noting that $d_i'Hd_i = 0$ for $i = 1, \ldots, k - 1$ shows that Part 2 holds true for $k \geq 2$. Part 2 holds true trivially for $k = 1$.

To show Part 3, since $d_i'Hd_j = 0$ for $i \neq j$, we get

\[
f(x_{k+1}) = f(x_1 + (x_{k+1} - x_1)) = f \left( x_1 + \sum_{j=1}^{k} \lambda_j d_j \right)
= f(x_1) + \nabla f(x_1)' \left( \sum_{j=1}^{k} \lambda_j d_j \right) + \frac{1}{2} \sum_{j=1}^{k} \lambda_j^2 d_i'Hd_j. \hspace{1cm} (8.26)
\]

Now suppose that $x - x_1 \in L(d_1, \ldots, d_k)$, so that $x$ can be written as $x_1 + \sum_{j=1}^{k} \mu_j d_j$. As in (8.26), we get

\[
f(x) = f(x_1) + \nabla f(x_1)' \left( \sum_{j=1}^{k} \mu_j d_j \right) + \frac{1}{2} \sum_{j=1}^{k} \mu_j^2 d_i'Hd_j. \hspace{1cm} (8.27)
\]

To complete the proof, we need to show that $f(x) \geq f(x_{k+1})$. By contradiction, suppose that $f(x) < f(x_{k+1})$. Then by (8.26) and (8.27), we must have
Unconstrained Optimization

By the definition of $\lambda_j$, note that $f(x_j + \lambda_j d_j) \leq f(x_j + \mu_j d_j)$ for each $j$. Therefore,

$$f(x_j) + \lambda_j \nabla f(x_j)'d_j + \frac{1}{2} \lambda_j^2 \mu_j d_j' H d_j \leq f(x_j) + \mu_j \nabla f(x_j)'d_j + \frac{1}{2} \mu_j^2 \mu_j d_j' H d_j.$$  

By Part 2, $\nabla f(x_j)'d_j = \nabla f(x_1)'d_j$, and substituting this in the inequality above, we get

$$\lambda_j \nabla f(x_1)'d_j + \frac{1}{2} \lambda_j^2 \mu_j d_j' H d_j \leq \mu_j \nabla f(x_1)'d_j + \frac{1}{2} \mu_j^2 \mu_j d_j' H d_j. \quad (8.29)$$

Summing (8.29) for $j = 1, ..., k$ contradicts (8.28). Thus, $x_{k+1}$ is a minimizing point over the manifold $x_1 + L(d_1, ..., d_k)$. In particular, since $d_1, ..., d_n$ are linearly independent, $L(d_1, ..., d_n) = R^n$, and hence, $x_{n+1}$ is a minimizing point of $f$ over $R^n$. This completes the proof.

Generating Conjugate Directions

In the remainder of this section we describe several methods for generating conjugate directions for quadratic forms. These methods lead naturally to powerful algorithms for minimizing both quadratic and nonquadratic functions. In particular, we discuss the classes of quasi-Newton and conjugate gradient methods.

Quasi-Newton Methods: Method of Davidon–Fletcher–Powell

This method was proposed by Davidon [1959] and later developed by Fletcher and Powell [1963]. The Davidon–Fletcher–Powell (DFP) method falls under the general class of quasi-Newton procedures, where the search directions are of the form $d_j = -D_j \nabla f(y)$, in lieu of $-H^{-1}(y)\nabla f(y)$, as in Newton's method. The negative gradient direction is thus deflected by premultiplying it by $-D_j$, where $D_j$ is an $n \times n$ positive definite symmetric matrix that approximates the inverse of the Hessian matrix. The positive definiteness property ensures that $d_j$ is a descent direction whenever $\nabla f(y) \neq 0$, since then, $d_j'\nabla f(y) < 0$. For the purpose of the next step, $D_{j+1}$ is formed by adding to $D_j$ two symmetric
matrices, each of rank one. Thus, this scheme is sometimes referred to as a rank-
two correction procedure. For quadratic functions, this update scheme is shown later to produce the exact representation of the actual inverse Hessian within \( n \) steps. The DFP process is also called a variable metric method because it can be interpreted as adopting the steepest descent step in the transformed space based on the Cholesky factorization of the positive definite matrix \( D_j \), as discussed in Section 8.7, where this transformation varies with \( D_j \) from iteration to iteration.

The quasi-Newton methods in which the quadratic approximation is permitted to be possibly indefinite are more generally called secant methods.

**Summary of the Davidon–Fletcher–Powell (DFP) Method**

We now summarize the Davidon–Fletcher–Powell (DFP) method for minimizing a differentiable function of several variables. In particular, if the function is quadratic, then, as shown later, the method yields conjugate directions and terminates in one complete iteration, that is, after searching along each of the conjugate directions as described below.

**Initialization Step**

Let \( \varepsilon > 0 \) be a termination tolerance. Choose an initial point \( x_1 \) and an initial symmetric positive definite matrix \( D_1 \). Let \( y_1 = x_1 \), let \( k = j = 1 \), and go to the Main Step.

**Main Step**

1. If \( \| \nabla f(y_j) \| < \varepsilon \), stop; otherwise, let \( d_j = -D_j \nabla f(y_j) \) and let \( \lambda_j \) be an optimal solution to the problem to minimize \( f(y_j + \lambda d_j) \) subject to \( \lambda \geq 0 \). Let \( y_{j+1} = y_j + \lambda_j d_j \). If \( j < n \), go to Step 2. If \( j = n \), let \( y_1 = x_{k+1} = y_{n+1} \), replace \( k \) by \( k + 1 \), let \( j = 1 \), and repeat Step 1.

2. Construct \( D_{j+1} \) as follows:

\[
D_{j+1} = D_j + \frac{p_j p_j' D_j q_j q_j'}{p_j' q_j},
\]

where

\[
p_j = \lambda_j d_j = y_{j+1} - y_j \tag{8.31}
\]

\[
q_j = \nabla f(y_{j+1}) - \nabla f(y_j). \tag{8.32}
\]

Replace \( j \) by \( j + 1 \), and go to Step 1.

We remark here that the inner loop of the foregoing algorithm resets the procedure every \( n \) steps (whenever \( j = n \) at Step 1). Any variant that resets every \( n' < n \) inner iteration steps is called a partial quasi-Newton method. This strategy
can be useful from the viewpoint of conserving storage when \( n' \ll n \), since then, the inverse Hessian approximation can be stored implicitly by, instead, storing only the generating vectors \( p_j \) and \( q_j \) themselves within the inner loop iterations.

### 8.8.4 Example

Consider the following problem:

\[
\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.
\]

A summary of the computations using the DFP method is given in Table 8.13. At each iteration, for \( j = 1, 2, \ldots, n \), \( d_j \) is given by \(-D_j \nabla f(y_j)\), where \( D_1 \) is the identity matrix and \( D_2 \) is computed from (8.30), (8.31), and (8.32). At Iteration \( k = 1 \), we have \( p_1 = (2.7, -1.49)' \) and \( q_1 = (44.73, -22.72)' \) in (8.30). At Iteration 2 we have \( p_1 = (-0.1, 0.05)' \) and \( q_1 = (-0.7, 0.8)' \), and finally, at Iteration 3 we have \( p_1 = (-0.02, 0.02)' \) and \( q_1 = (-0.14, 0.24)' \). The point \( y_{j+1} \) is computed by optimizing along the direction \( d_j \) starting from \( y_j \) for \( j = 1, 2 \). The procedure is terminated at the point \( y_2 = (2.115, 1.058)' \) in the fourth iteration, since \( \| \nabla f(y_2) \| = 0.006 \) is quite small. The path taken by the method is depicted in Figure 8.22.

Lemma 8.8.5 shows that each matrix \( D_j \) is positive definite and \( d_j \) is a direction of descent.

### 8.8.5 Lemma

Let \( y_1 \in \mathbb{R}^n \), and let \( D_1 \) be an initial positive definite symmetric matrix. For \( j = 1, \ldots, n \), let \( y_{j+1} = y_j + \lambda_j d_j \), where \( d_j = -D_j \nabla f(y_j) \), and \( \lambda_j \) solves the problem to minimize \( f(y_j + \lambda d_j) \) subject to \( \lambda \geq 0 \). Furthermore, for \( j = 1, \ldots, n - 1 \), let \( D_{j+1} \) be given by (8.30), (8.31), and (8.32). If \( \nabla f(y_j) \neq 0 \) for \( j = 1, \ldots, n \), \( D_1, \ldots, D_n \) are symmetric and positive definite so that \( d_1, \ldots, d_n \) are descent directions.

**Proof**

We prove the result by induction. For \( j = 1 \), \( D_1 \) is symmetric and positive definite by assumption. Furthermore, \( \nabla f(y_1)'d_1 = -\nabla f(y_1)'D_1\nabla f(y_1) < 0 \), since \( D_1 \) is positive definite. By Theorem 4.1.2, \( d_1 \) is a descent direction. We
Table 8.13 Summary of Computations for the Davidon–Fletcher–Powell Method

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$x_k$</th>
<th>$y_j$</th>
<th>$\nabla f(y_j)$</th>
<th>$|\nabla f(y_j)|$</th>
<th>$D_j$</th>
<th>$d_j$</th>
<th>$\lambda_j$</th>
<th>$y_{j+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.00, 3.00)</td>
<td>(0.00, 3.00)</td>
<td>(-44.00, 24.00)</td>
<td>50.12</td>
<td>[1 0]</td>
<td>(44.00, -24.00)</td>
<td>0.062</td>
<td>(2.70, 1.51)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(2.70, 1.51)</td>
<td>(0.73, 1.28)</td>
<td>0.34</td>
<td>0.99</td>
<td>[0.25 0.38]</td>
<td>(-0.67, -1.31)</td>
<td>0.22</td>
<td>(2.55, 1.22)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(2.55, 1.22)</td>
<td>(0.89, -0.44)</td>
<td>0.40</td>
<td>0.99</td>
<td>[0.65 0.45]</td>
<td>(-0.28, -0.25)</td>
<td>0.64</td>
<td>(2.27, 1.11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(2.27, 1.11)</td>
<td>(0.18, -0.20)</td>
<td>0.27</td>
<td>0.99</td>
<td>[0.80 0.38]</td>
<td>(-0.05, -0.03)</td>
<td>2.64</td>
<td>(2.12, 1.05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.12, 1.05)</td>
<td>(0.05, -0.08)</td>
<td>0.09</td>
<td>0.99</td>
<td>[1 0]</td>
<td>(-0.05, 0.08)</td>
<td>0.10</td>
<td>(2.115, 1.058)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(2.115, 1.058)</td>
<td>(0.004, 0.004)</td>
<td>0.006</td>
<td>0.99</td>
<td>[1 0]</td>
<td>(-0.05, 0.08)</td>
<td>0.10</td>
<td>(2.115, 1.058)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
shall assume that the result holds true for $j \leq n - 1$ and then show that it holds for $j + 1$. Let $x$ be a nonzero vector in $\mathbb{R}^n$; then, by (8.30), we have

$$\begin{equation}
x'^jD_{j+1}x = x'^jD_jx + \frac{(x'^j p_j)^2}{p'_jq_j} - \frac{(x'^j D_j q_j)^2}{q'_jq_j}.
\end{equation}$$

Since $D_j$ is a symmetric positive definite matrix, there exists a positive definite symmetric matrix $D_j^{1/2}$ such that $D_j = D_j^{1/2}D_j^{1/2}$. Let $a = D_j^{1/2}x$ and $b = D_j^{1/2}q_j$. Then $x'^jD_jx = a'^ja$, $q'_jq_j = b'b$, and $x'^jD_jq_j = a'b$. Substituting in (8.33), we get

$$\begin{equation}
x'^jD_{j+1}x = \frac{(a'^ja)(b'b) - (a'b)^2}{b'b} + \frac{(x'^j p_j)^2}{p'_jq_j}.
\end{equation}$$

Figure 8.22 Davidon–Fletcher–Powell method.
By the Schwartz inequality, \((a' a)(b' b) \geq (a' b)^2\). Thus, to show that \(x' D_{j+1} x \geq 0\), it suffices to show that \(p_j' q_j > 0\) and that \(b' b > 0\). From (8.31) and (8.32) it follows that
\[
p_j' q_j = \lambda_j d_j' [\nabla f(y_{j+1}) - \nabla f(y_j)].
\]
The reader may note that \(d_j' \nabla f(y_{j+1}) = 0\), and by definition, \(d_j = -D_j \nabla f(y_j)\). Substituting these in the above equation, it follows that
\[
p_j' q_j = \lambda_j \nabla f(y_j)' D_j \nabla f(y_j).
\] (8.35)

Note that \(\nabla f(y_j) \neq 0\) by assumption, and that \(D_j\) is positive definite, so that \(\nabla f(y_j)' D_j \nabla f(y_j) > 0\). Furthermore, \(d_j\) is a descent direction and, hence, \(\lambda_j > 0\). Therefore, from (8.35), \(p_j' q_j > 0\). Furthermore, \(q_j \neq 0\) and, hence, \(b' b = q_j' D_j q_j > 0\).

We now show that \(x' D_{j+1} x > 0\). By contradiction, suppose that \(x' D_{j+1} x = 0\). This is possible only if \((a' a)(b' b) = (a' b)^2\) and \(p_j' x = 0\). First, note that \((a' a)(b' b) = (a' b)^2\) only if \(a = \lambda b\); that is, \(D_j^{1/2} x = \lambda D_j^{1/2} q_j\). Thus, \(x = \lambda q_j\). Since \(x \neq 0\), we have \(\lambda \neq 0\). Now \(0 = p_j' x = \lambda p_j' q_j\) contradicts the fact that \(p_j' q_j > 0\) and \(\lambda \neq 0\). Therefore, \(x' D_{j+1} x > 0\), so that \(D_{j+1}\) is positive definite.

Since \(\nabla f(y_{j+1}) \neq 0\) and since \(D_{j+1}\) is positive definite, \(\nabla f(y_{j+1})' d_{j+1} = -\nabla f(y_{j+1})' D_{j+1} \nabla f(y_{j+1}) < 0\). By Theorem 4.1.2, then, \(d_{j+1}\) is a descent direction. This completes the proof.

**Quadratic Case**

If the objective function \(f\) is quadratic, then by Theorem 8.8.6, the directions \(d_1, \ldots, d_n\) generated by the DFP method are conjugate. Therefore, by Part 3 of Theorem 8.8.3, the method stops after one complete iteration with an optimal solution. Furthermore, the matrix \(D_{n+1}\) obtained at the end of the iteration is precisely the inverse of the Hessian matrix \(H\).

**8.8.6 Theorem**

Let \(H\) be an \(n \times n\) symmetric positive definite matrix, and consider the problem to minimize \(f(x) = c' x + (1/2)x' H x\) subject to \(x \in \mathbb{R}^n\). Suppose that the
problem is solved by the DFP method, starting with an initial point $y_1$ and a
symmetric positive definite matrix $D_1$. In particular, for $j = 1, \ldots, n$, let $\lambda_j$ be an
optimal solution to the problem to minimize $f(y_j + \lambda d_j)$ subject to $\lambda \geq 0$, and
let $y_{j+1} = y_j + \lambda_j d_j$, where $d_j = -D_j \nabla f(y_j)$ and $D_j$ is determined by
(8.30), (8.31), and (8.32). If $\nabla f(y_j) \neq 0$ for each $j$, then the directions $d_1, \ldots, d_n$
are $H$-conjugate and $D_{n+1} = H^{-1}$. Furthermore, $y_{n+1}$ is an optimal solution to
the problem.

**Proof**

We first show that for any $j$ with $1 \leq j \leq n$, we must have the following
conditions:

1. $d_1, \ldots, d_j$ are linearly independent.
2. $d_i^T H d_k = 0$ for $i \neq k; i, k \leq j$. (8.36)
3. $D_{j+1} H p_k = p_k$ or, equivalently, $D_{j+1} H d_k = d_k$ for $1 \leq k \leq j$,
   where $p_k = \lambda_k d_k$.

We prove this result by induction. For $j = 1$, parts 1 and 2 are obvious. To
prove Part 3, first note that for any $k$,

$$H p_k = H(\lambda_k d_k) = H(y_{k+1} - y_k) = \nabla f(y_{k+1}) - \nabla f(y_k) = q_k.$$ (8.37)

In particular, $H p_1 = q_1$. Thus, letting $j = 1$ in (8.30), we get

$$D_2 H p_1 = \left( D_1 + \frac{p_1 p_1^T}{p_1^T q_1} - \frac{D_1 q_1 q_1^T D_1}{q_1^T D_1 q_1} \right) q_1 = p_1$$

so that Part 3 holds true for $j = 1$.

Now suppose that Parts 1, 2, and 3 hold true for $j \leq n - 1$. To show that
they also hold true for $j + 1$, first recall by Part 1 of Theorem 8.8.3 that
$d_i^T \nabla f(y_{j+1}) = 0$ for $i \leq j$. By the induction hypothesis of Part 3, $d_i = D_{j+1} H d_i$
for $i \leq j$. Thus, for $i \leq j$ we have

$$0 = d_i^T \nabla f(y_{j+1}) + d_i^T H d_{j+1} \nabla f(y_{j+1}) = -d_i^T H d_{j+1}.$$

In view of the induction hypothesis for Part 2, the above equation shows that
Part 2 also holds true for $j + 1$.

Now, we show that Part 3 holds true for $j + 1$. Letting $k \leq j + 1$ yields

$$D_{j+2} H p_k = \left( D_{j+1} + \frac{p_{j+1} p_{j+1}^T}{p_{j+1}^T q_{j+1}} - \frac{D_{j+1} q_{j+1} q_{j+1}^T D_{j+1}}{q_{j+1}^T D_{j+1} q_{j+1}} \right) H p_k.$$ (8.38)
Noting (8.37) and letting \( k = j + 1 \) in (8.38), it follows that \( D_{j+2}H_{p_{j+1}} = p_{j+1} \). Now let \( k \leq j \). Since Part 2 holds true for \( j + 1 \),

\[
p'_{j+1}H_{p_k} = \lambda_k \lambda_{j+1}d'_{j+1}H_{d_k} = 0. \tag{8.39}
\]

Noting the induction hypothesis for Part 3, (8.37), and the fact that Part 2 holds true for \( j + 1 \), we get

\[
q'_{j+1}D_{j+1}H_{p_k} = q'_{j+1}p_k = p'_{j+1}H_{p_k} = \lambda_{j+1} \lambda_k d'_{j+1}H_{d_k} = 0. \tag{8.40}
\]

Substituting (8.39) and (8.40) in (8.38), and noting the induction hypothesis for Part 3, we get

\[
D_{j+2}H_{p_k} = D_{j+1}H_{p_k} = p_k.
\]

Thus, Part 3 holds true for \( j + 1 \).

To complete the induction argument, we only need to show that Part 1 holds true for \( j + 1 \). Suppose that \( \sum_{i=1}^{j+1} \alpha_i d_i = 0 \). Multiplying by \( d'_{j+1}H \) and noting that Part 2 holds true for \( j + 1 \), it follows that \( \alpha_{j+1}d'_{j+1}Hd_{j+1} = 0 \). By assumption, \( \nabla f(y_{j+1}) \neq 0 \), and by Lemma 8.8.5, \( D_{j+1} \) is positive definite, so that \( d_{j+1} = -D_{j+1}\nabla f(y_{j+1}) \neq 0 \). Since \( H \) is positive definite, \( d'_{j+1}Hd_{j+1} \neq 0 \), and hence, \( \alpha_{j+1} = 0 \). This in turn implies that \( \sum_{i=1}^{j} \alpha_i d_i = 0 \); and since \( d_1, \ldots, d_j \) are linearly independent by the induction hypothesis, \( \alpha_i = 0 \) for \( i = 1, \ldots, j \). Thus, \( d_1, \ldots, d_{j+1} \) are linearly independent and Part 1 holds true for \( j + 1 \). Thus, Parts 1, 2, and 3 hold true. In particular, the conjugacy of \( d_1, \ldots, d_n \) follows from Parts 1 and 2 by letting \( j = n \).

Now, let \( j = n \) in Part 3. Then \( D_{n+1}H_{d_k} = d_k \) for \( k = 1, \ldots, n \). If we let \( D \) be the matrix whose columns are \( d_1, \ldots, d_n \), then \( D_{n+1}HD = D \). Since \( D \) is invertible, \( D_{n+1}H = I \), which is possible only if \( D_{n+1} = H^{-1} \). Finally, \( y_{n+1} \) is an optimal solution by Theorem 8.8.3.

**Insightful Derivation of the DFP Method**

At each step of the DFP method we have seen that given some approximation \( D_j \) to the inverse Hessian matrix, we computed the search direction \( d_j = -D_j\nabla f(y_j) \) by deflecting the negative gradient of \( f \) at the current solution \( y_j \) using this approximation \( D_j \) in the spirit of Newton's method. We then performed a line search along this direction, and based on the resulting solution \( y_{j+1} \) and the gradient \( \nabla f(y_{j+1}) \) at this point, we obtained an updated approximation \( D_{j+1} \) according to (8.30), (8.31), and (8.32). As seen in Theorem
8.8.6, if \( f \) is a quadratic function given by \( f(x) = c^T x + \frac{1}{2} x^T H x; \ x \in \mathbb{R}^n \), where \( H \) is symmetric and positive definite; and if \( \nabla f(y_j) \neq 0, j = 1, \ldots, n \), then we indeed obtain \( D_{n+1} = H^{-1} \). In fact, observe from Parts 1 and 3 of Theorem 8.8.6 that for each \( j \in \{1, \ldots, n\} \), the vectors \( p_1, \ldots, p_j \) are linearly independent eigenvectors of \( D_{j+1}H \) having eigenvalues equal to 1. Hence, at each step of the method, the revised approximation accumulates one additional linearly independent eigenvector, with a unit eigenvalue for the product \( D_{j+1}H \), until \( D_{n+1}H \) finally has all its \( n \) eigenvalues equal to 1, giving \( D_{n+1}HP = P \), where \( P \) is the nonsingular matrix of the eigenvectors of \( D_{n+1}H \). Hence, \( D_{n+1}H = I \), or \( D_{n+1} = H^{-1} \).

Based on the foregoing observation, let us derive the update scheme (8.30) for the DFP method and use this derivation to motivate other more prominent updates. Toward this end, suppose that we have some symmetric, positive definite approximation \( D_j \) of the inverse Hessian matrix for which \( p_1, \ldots, p_{j-1} \) are the eigenvectors of \( D_jH \) with unit eigenvalues. (For \( j = 1, \) no such vector exists.) Adopting the inductive scheme of Theorem 8.8.6, assume that these eigenvectors are linearly independent and are \( H \)-conjugate. Now, given the current point \( y_j \), we conduct a line search along the direction \( d_j = -D_j \nabla f(y_j) \) to obtain the new point \( y_{j+1} \) and, accordingly, we define

\[
\begin{align*}
   p_j &= (y_{j+1} - y_j) \\
   q_j &= \nabla f(y_{j+1}) - \nabla f(y_j) = H(y_{j+1} - y_j) = Hp_j.
\end{align*}
\]

Following the argument in the proof of Theorem 8.8.6, the vectors \( p_k = \lambda_k d_k, k = 1, \ldots, j \), are easily shown to be linearly independent and \( H \)-conjugate. We now want to construct a matrix

\[
D_{j+1} = D_j + C_j,
\]

where \( C_j \) is some symmetric correction matrix, which ensures that \( p_1, \ldots, p_j \) are eigenvectors of \( D_{j+1}H \) having unit eigenvalues. Hence, we want \( D_{j+1}HP_k = p_k \) or, from (8.41), that \( D_{j+1}q_k = p_k \) for \( k = 1, \ldots, j \). For \( 1 \leq k < j \), this translates to requiring that \( p_k = D_jq_k + C_jq_k = D_jHP_k + C_jq_k = p_k + C_jq_k \), or that

\[
C_jq_k = 0 \quad \text{for} \ k = 1, \ldots, j - 1.
\]

For \( k = j \), the aforementioned condition

\[
D_{j+1}q_j = p_j
\]
is called the quasi-Newton condition, or the secant equation, the latter term leading to the alternative name secant updates for this type of scheme. This condition translates to the requirement that

\[ C_j q_j = p_j - D_j q_j. \] (8.44)

Now if \( C_j \) had a symmetric rank-one term \( p_j p'_j / q'_j q_j \), then \( C_j q_j \) operating on this term would yield \( p_j \), as required in (8.44). Similarly, if \( C_j \) had a symmetric rank-one term, \( -(D_j q_j)(D_j q_j)'/(D_j q_j)' q_j \), then \( C_j q_j \) operating on this term would yield \( -D_j q_j \), as required in (8.44). This therefore leads to the rank-two DFP update (8.30) via the correction term,

\[ C_j = \frac{p'_j p'_j - D_j q_j q'_j D_j}{q'_j D_j q_j} = C_j^{DFP}, \] (8.45)

which satisfies the quasi-Newton condition (8.43) via (8.44). (Note that as in Lemma 8.8.5, \( D_{j+1} = D_j + C_j \) is symmetric and positive definite.) Moreover, (8.42) also holds since for any \( k \in \{1, \ldots, j-1\} \), we have from (8.45) and (8.41) that

\[ C_j q_k = C_j H p_k = \frac{p'_j p'_j H p_k}{p'_j q_j} - \frac{D_j q_j p'_j H D_j H p_k}{q'_j D_j q_j} = 0 \]

since \( p'_j H p_k = 0 \) in the first term and \( p'_j H D_j H p_k = p'_j H p_k = 0 \) in the second term as well. Hence, following this sequence of corrections, we shall ultimately obtain \( D_{n+1} H = I \) or \( D_{n+1} = H^{-1} \).

**Broyden Family and Broyden–Fletcher–Goldfarb–Shanno (BFGS) Updates**

The reader might have observed in the foregoing derivation of \( C_j^{DFP} \) that there was a degree of flexibility in prescribing the correction matrix \( C_j \), the restriction being to satisfy the quasi-Newton condition (8.44) along with (8.42) and to maintain symmetry and positive definiteness of \( D_{j+1} = D_j + C_j \). In light of this, the Broyden updates suggest the use of the correction matrix \( C_j = C_j^B \) given by the following family parameterized by \( \phi \):

\[ C_j^B = C_j^{DFP} + \frac{\phi r_j v'_j y'_j}{p'_j q_j}, \] (8.46)
where \( v_j = p_j - (1/\tau_j)D_jq_j \) and where \( \tau_j \) is chosen so that the quasi-Newton condition (8.44) holds by virtue of \( v_j'q_j \) being zero. This implies that \([p_j - D_jq_j/\tau_j]q_j = 0\), or that

\[
\tau_j = \frac{q_j'D_jq_j}{p_j'q_j} > 0. \tag{8.47}
\]

Note that for \( 1 \leq k < j \), we have

\[
v_j'q_k = p_j'q_k - \frac{1}{\tau_j} q_j'D_jq_k = p_j'Hp_k - \frac{1}{\tau_j} p_j'D_jHq_k = 0
\]

because \( p_j'Hp_k = 0 \) by conjugacy and \( p_j'[D_jHq_k] = p_j'Hq_k = 0 \), since \( p_k \) is an eigenvector of \( D_jH \) having a unit eigenvalue. Hence, (8.42) also continues to hold true. Moreover, it is clear that \( D_{j+1} = D_j + C_j^B \) continues to be symmetric and, at least for \( \phi \geq 0 \), positive definite. Hence, the correction matrix (8.46)–(8.47) in this case yields a valid sequence of updates satisfying the assertion of Theorem 8.8.6.

For the value \( \phi = 1 \), the Broyden family yields a very useful special case, which coincides with that derived independently by Broyden, Fletcher, Goldfarb, and Shanno. This update, known as the BFGS update, or the positive definite secant update, has been consistently shown in many computational studies to dominate other updating schemes in its overall performance. In contrast, the DFP update has been observed to exhibit numerical difficulties, sometimes having the tendency to produce near-singular Hessian approximations. The additional correction term in (8.46) seems to alleviate this propensity.

To derive this update correction \( C_j^{BFGS} \), say, we simply substitute (8.47) into (8.46), and simplify (8.46) using \( \phi = 1 \) to get

\[
C_j^{BFGS} = C_j^B (\phi = 1) = \frac{p_jb_j'}{p_j'q_j} \left( 1 + \frac{q_j'D_jq_j}{p_j'q_j} \right) D_jq_j + \frac{p_jq_j'D_j}{p_j'q_j}. \tag{8.48}
\]

Since with \( \phi = 0 \) we have \( C_j^B = C_j^{DFP} \), we can write (8.46) as

\[
C_j^B = (1 - \phi)C_j^{DFP} + \phi C_j^{BFGS}. \tag{8.49}
\]

The above discussion assumes the use of a constant value \( \phi \) in (8.46). This is known as a pure Broyden update. However, for the analytical results to hold true, it is not necessary to work with a constant value of \( \phi \). A variable value \( \phi_j \) can be chosen from one iteration to the next if so desired. However, there is a...
value of $\phi$ in (8.46) that will make $d_{j+1} = -D_{j+1} \nabla f(y_{j+1})$ identically zero (see Exercise 8.35), namely,

$$\phi = -\frac{[\nabla f(y_j)^t D_j \nabla f(y_j)] q_j^t D_j q_j}{\nabla f(y_{j+1})^t D_j \nabla f(y_{j+1})}.$$  

(8.50)

Hence, the algorithm stalls and, in particular, $D_{j+1}$ becomes singular and loses positive definiteness. Such a value of $\phi$ is said to be degenerate, and should be avoided. For this reason, as a safeguard, $\phi$ is usually taken to be nonnegative, although sometimes, admitting negative values seems to be computationally attractive. In this connection, note that for a general differentiable function, if perfect line searches are performed (i.e., either an exact minimum, or in the nonconvex case, the first local minimum along a search direction is found), then it can be shown that the sequence of iterates generated by the Broyden family is invariant with respect to the choice of the parameter $\phi$ as long as nondegenerate $\phi$ values are chosen (see the Notes and References section). Hence, the choice of $\phi$ becomes critical only with inexact line searches. Also, if inaccurate line searches are used, then maintaining the positive definiteness of the Hessian approximations becomes a matter of concern. In particular, this motivates the following strategy.

**Updating Hessian Approximations**

Note that in a spirit similar to the foregoing derivations, we could alternatively have started with a symmetric, positive definite approximation $B_1$ to the Hessian $H$ itself, and then updated this to produce a sequence of symmetric, positive definite approximations according to $B_{j+1} = B_j + C_j$ for $j = 1, \ldots, n$.

Again, for each $j = 1, \ldots, n$, we would like $p_1, \ldots, p_j$ to be eigenvectors of $H^{-1}B_{j+1}$ having eigenvalues of 1, so that for $j = n$ we would obtain $H^{-1}B_{n+1} = I$ or $B_{n+1} = H$ itself. Proceeding inductively as before, given that $p_1, \ldots, p_{j-1}$ are eigenvectors of $H^{-1}B_j$ associated with unit eigenvalues, we need to construct a correction matrix $\bar{C}_j$ such that $H^{-1}(B_j + \bar{C}_j)p_k = p_k$ for $k = 1, \ldots, j$. In other words, multiplying throughout by $H$ and noting that $q_k = H p_k$ for $k = 1, \ldots, j$ by (8.41), if we are given that

$$B_j p_k = q_k \quad \text{for } k = 1, \ldots, j - 1$$  

(8.51)

we are required to ensure that $(B_j + \bar{C}_j)p_k = q_k$ for $k = 1, \ldots, j$ or, using (8.51), that

$$\bar{C}_j p_k = 0 \quad \text{for } 1 \leq k \leq j - 1 \quad \text{and} \quad \bar{C}_j p_j = q_j - B_j p_j.$$  

(8.52)
Comparing (8.51) with the condition $D_j q_k = p_k$ for $k = 1, \ldots, j - 1$ and, similarly, comparing (8.52) with (8.42) and (8.44), we observe that the present analysis differs from the foregoing analysis involving an update of inverse Hessians in that the role of $D_j$ and $B_j$, and that of $p_j$ and $q_j$, are interchanged. By symmetry, we can derive a formula for $\overline{C}_j$ simply by replacing $D_j$ by $B_j$ and by interchanging $p_j$ and $q_j$ in (8.45). An update obtained in this fashion is called a complementary update, or dual update, to the preceding one. Of course, the dual of the dual formula will naturally yield the original formula. The $\overline{C}_j$ derived as the dual to $C_j^{DFP}$ was actually obtained independently by Broyden, Fletcher, Goldfarb, and Shanno in 1970, and the update is therefore known as the BFGS update. Hence, we have

$$\overline{C}_j^{BFGS} = \frac{q_j q_j' - B_j p_j p_j'}{q_j' p_j} B_j.$$  
  \hspace{1cm} (8.53)

In Exercise 8.37 we ask the reader to derive (8.53) directly following the derivation of (8.41) through (8.45).

Note that the relationship between $\overline{C}_j^{BFGS}$ and $C_j^{BFGS}$ is as follows:

$$D_{j+1} = D_j + C_j^{BFGS} = B_j^{-1} = (B_j + C_j^{BFGS})^{-1}. \hspace{1cm} (8.54)$$

That is, $D_{j+1} q_k = p_k$ for $k = 1, \ldots, j$ implies that $D_j^{-1} p_k = q_k$ or that $B_{j+1} = D_{j+1}^{-1}$ indeed satisfies (8.51) (written for $j + 1$). In fact, the inverse relationship (8.54) between (8.48) and (8.53) can readily be verified (see Exercise 8.36) by using two sequential applications of the Sherman–Morrison–Woodbury formula given below, which is valid for any general $n \times n$ matrix $A$ and $n \times 1$ vectors $a$ and $b$, given that the inverse exists (or equivalently, given that $1 + b' A^{-1} a \neq 0$):

$$(A + ab')^{-1} = A^{-1} - \frac{A^{-1} ab' A^{-1}}{1 + b' A^{-1} a}. \hspace{1cm} (8.55)$$

Note that if the Hessian approximations $B_j$ are generated as above, then the search direction $d_j$ at any step needs to be obtained by solving the system of equations $B_j d_j = -\nabla f(y_j)$. This can be done more conveniently by maintaining a Cholesky factorization $\mathcal{L}_j D_j \mathcal{L}_j'$ of $B_j$, where $\mathcal{L}_j$ is a lower triangular matrix and $\mathcal{D}_j$ is a diagonal matrix. Besides the numerical benefits of adopting this procedure, it can also be helpful in that the condition number of $\mathcal{D}_j$ can be useful in assessing the ill-conditioning status of $B_j$, and the positive
definiteness of $B_j$ can be verified by checking the positivity of the diagonal elements of $D_j$. Hence, when an update of $D_j$ reveals a loss of positive definiteness, alternative steps can be taken by restoring the diagonal elements of $D_{j+1}$ to be positive.

**Scaling of Quasi-Newton Algorithms**

Let us conclude our discussion on quasi-Newton methods by making a brief but important comment on adopting a proper scaling of the updates generated by these methods. In our discussion leading to the derivation of (8.41)–(8.45), we learned that at each step $j$, the revised update $D_{j+1}$ had an additional eigenvector associated with a unit eigenvalue for the matrix $D_{j+1\cdot}$. Hence, if, for example, $D_1$ is chosen such that the eigenvalues of $D_1\cdot$ are all significantly larger than unity, then since these eigenvalues are transformed to unity one at a time as the algorithm proceeds, one can expect an unfavorable ratio of the largest to smallest eigenvalues of $D_{j\cdot}$ at the intermediate steps.

When minimizing nonquadratic functions and/or employing inexact line searches, in particular, such a phenomenon can result in ill-conditioning effects and exhibit poor convergence performance. To alleviate this, it is useful to multiply each $D_j$ by some scale factor $s_j > 0$ before using the update formula. With exact line searches, this can be shown to preserve the conjugacy property in the quadratic case, although we may no longer have $D_{n+1} = H^{-1}$. However, the focus here is to improve the single-step rather than the $n$-step convergence behavior of the algorithm. Methods that automatically prescribe scale factors in a manner such that if the function is quadratic, then the eigenvalues of $s_jD_j\cdot$ tend to be spread above and below unity are called *self-scaling methods*. We refer the reader to the Notes and References section for further reading on this subject.

**Conjugate Gradient Methods**

Conjugate gradient methods were proposed by Hestenes and Stiefel in 1952 for solving systems of linear equations. The use of this method for unconstrained optimization was prompted by the fact that the minimization of a positive definite quadratic function is equivalent to solving the linear equation system that results when its gradient is set at zero. Actually, conjugate gradient methods were first extended to solving nonlinear equation systems and general unconstrained minimization problems by Fletcher and Reeves in 1964. Although these methods are typically less efficient and less robust than quasi-Newton methods, they have very modest storage requirements (only three $n$-vectors are required for the method of Fletcher and Reeves described below) and are quite indispensable for large problems ($n$ exceeding about 100) when quasi-Newton methods become impractical because of the size of the Hessian matrix. Some
very successful applications are reported by Fletcher [1987] in the context of atomic structures, where problems having 3000 variables were solved using only about 50 gradient evaluations, and by Reid [1971], who solved some linear partial differential equations having some 4000 variables in about 40 iterations. Moreover, conjugate gradient methods have the advantage of simplicity, being gradient deflection methods that deflect the negative gradient direction using the previous direction. This deflection can alternatively be viewed as an update of a fixed, symmetric, positive definite matrix, usually the identity matrix, in the spirit of quasi-Newton methods. For this reason they are sometimes referred to as fixed-metric methods in contrast to the term variable-metric methods, which applies to quasi-Newton procedures. Again, these are conjugate direction methods that converge in, at most, \( n \) iterations for unconstrained quadratic optimization problems in \( \mathbb{R}^n \) when using exact line searches. In fact, for the latter case, they generate directions identical to the BFGS method, as shown later.

The basic approach of conjugate gradient methods for minimizing a differentiable function \( f : \mathbb{R}^n \to \mathbb{R} \) is to generate a sequence of iterates \( y_j \) according to

\[
y_{j+1} = y_j + \lambda_j d_j
\]

(8.56a)

where \( d_j \) is the search direction and \( \lambda_j \) is the step length that minimizes \( f \) along \( d_j \) from the point \( y_j \). For \( j = 1 \), the search direction \( d_1 = -\nabla f(y_1) \) can be used, and for subsequent iterations, given \( y_{j+1} \) with \( \nabla f(y_{j+1}) \neq 0 \) for \( j \geq 1 \), we use

\[
d_{j+1} = -\nabla f(y_{j+1}) + \alpha_j d_j,
\]

(8.56b)

where \( \alpha_j \) is a suitable deflection parameter that characterizes a particular conjugate gradient method. Note that we can write \( d_{j+1} \) in (8.56b) whenever \( \alpha_j \geq 0 \) as

\[
d_{j+1} = \frac{1}{\mu} [\mu \nabla f(y_{j+1}) + (1 - \mu)d_j],
\]

where \( \mu = 1/(1 + \alpha_j) \), so \( d_{j+1} \) can then be essentially viewed as a convex combination of the current steepest descent direction and the direction used at the last iteration.

Now suppose that we assume \( f \) to be a quadratic function having a positive definite Hessian \( H \), and that we require \( d_{j+1} \) and \( d_j \) to be \( H \)-conjugate. From (8.56a) and (8.41), \( d_{j+1}'Hd_j = 0 \) amounts to requiring that \( 0 = d_{j+1}'Hp_j = \)
Using this in (8.56b) gives Hestenes and Stiefel's [1952] choice for \( \alpha_j \), used even in nonquadratic situations by assuming a local quadratic behavior, as

\[
\alpha_j^{HS} = \frac{\nabla f(y_{j+1})^T q_j}{d_j^T q_j} = \frac{\lambda_j \nabla f(y_{j+1})^T q_j}{p_j^T q_j}. 
\]  

(8.57)

When exact line searches are performed, we have \( d_j^T \nabla f(y_{j+1}) = 0 = d_{j-1}^T \nabla f(y_j) \), leading to

\[
d_j^T q_j = -d_j^T \nabla f(y_j) = [\nabla f(y_j) - \alpha_{j-1} d_{j-1}]^T \nabla f(y_j) = \| \nabla f(y_j) \|^2. \]

Substituting this into (8.57) yields Polak and Ribiere's [1969] choice for \( \alpha_j \) as

\[
\alpha_j^{PR} = \frac{\nabla f(y_{j+1})^T q_j}{\| \nabla f(y_j) \|^2}. 
\]  

(8.58)

Furthermore, if \( f \) is quadratic and if exact line searches are performed, we have, using (8.56) along with \( \nabla f(y_{j+1})^T d_j = 0 = \nabla f(y_j)^T d_{j-1} \) as above, that

\[
\nabla f(y_{j+1})^T \nabla f(y_j) = \nabla f(y_{j+1})^T [\alpha_{j-1} d_{j-1} - d_j] \\
= \alpha_{j-1} \nabla f(y_{j+1})^T d_{j-1} = \alpha_{j-1} [\nabla f(y_j) + \lambda_j H d_{j-1}]^T d_{j-1} \\
= \alpha_{j-1} \lambda_j d_j^T H d_{j-1} = 0
\]

by the H-conjugacy of \( d_j \) and \( d_{j-1} \) (where \( d_0 = 0 \)). Hence,

\[
\nabla f(y_{j+1})^T \nabla f(y_j) = 0. 
\]  

(8.59)

Substituting this into (8.58) and using (8.41) gives Fletcher and Reeves's [1964] choice of \( \alpha_j \) as

\[
\alpha_j^{FR} = \frac{\| \nabla f(y_{j+1}) \|^2}{\| \nabla f(y_j) \|^2}. 
\]  

(8.60)

We now proceed to present and formally analyze the conjugate gradient method using Fletcher and Reeves's choice (8.60) for \( \alpha_j \). A similar discussion follows for other choices as well.

**Summary of the Conjugate Gradient Method of Fletcher and Reeves**

A summary of this conjugate gradient method for minimizing a general differentiable function is given below.
Initialization Step
Choose a termination scalar $\varepsilon > 0$ and an initial point $x_0$. Let $y_1 = x_0$, $d_1 = -\nabla f(y_1)$, $k = j = 1$, and go to the Main Step.

Main Step

1. If $\|\nabla f(y_j)\| < \varepsilon$, stop. Otherwise, let $\lambda_j$ be an optimal solution to the problem to minimize $f(y_j + \lambda d_j)$ subject to $\lambda \geq 0$, and let $y_{j+1} = y_j + \lambda_j d_j$. If $j < n$, go to Step 2; otherwise, go to Step 3.

2. Let $d_{j+1} = -\nabla f(y_{j+1}) + \alpha_j d_j$, where

$$
\alpha_j = \frac{\|\nabla f(y_{j+1})\|^2}{\|\nabla f(y_j)\|^2}.
$$

Replace $j$ by $j + 1$, and go to Step 1.

3. Let $y_1 = x_{k+1} = y_{n+1}$, and let $d_1 = -\nabla f(y_1)$. Let $j = 1$, replace $k$ by $k + 1$, and go to Step 1.

8.8.7 Example
Consider the following problem:

$$
\text{Minimize } (x_1 - 2)^4 + (x_1 - 2x_2)^2.
$$

The summary of the computations using the method of Fletcher and Reeves is given in Table 8.14. At each iteration, $d_1$ is given by $-\nabla f(y_1)$, and $d_2$ is given by $d_2 = -\nabla f(y_2) + \alpha_1 d_1$, where $\alpha_1 = \|\nabla f(y_2)\|^2 / \|\nabla f(y_1)\|^2$. Furthermore, $y_{j+1}$ is obtained by optimizing along $d_j$, starting from $y_j$. At Iteration 4, the point $y_2 = (2.185, 1.094)^T$, which is very close to the optimal point $(2.00, 1.00)$, is reached. Since the norm of the gradient is equal to 0.02, which is, say, sufficiently small, we stop here. The progress of the algorithm is shown in Figure 8.23.

Quadratic Case
If the function $f$ is quadratic, Theorem 8.8.8 shows that the directions $d_1, \ldots, d_n$ generated are indeed conjugate, and hence, by Theorem 8.8.3, the conjugate gradient algorithm produces an optimal solution after one complete application of the Main Step, that is, after at most $n$ line searches have been performed.
Table 8.14 Summary of Computations for the Method of Fletcher and Reeves

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$x_k$</th>
<th>$y_j$</th>
<th>$\nabla f(y_j)$</th>
<th>$|\nabla f(y_j)|$</th>
<th>$\alpha_{j-1}$</th>
<th>$d_j$</th>
<th>$\lambda_j$</th>
<th>$y_{j+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$f(x_k)$</td>
<td>$f(y_j)$</td>
<td>(-44.00, 24.00)</td>
<td>50.12</td>
<td>—</td>
<td>(44.00, -24.00)</td>
<td>0.062</td>
<td>(2.70, 1.51)</td>
</tr>
<tr>
<td>1</td>
<td>(0.00, 3.00)</td>
<td>52.00</td>
<td>1</td>
<td>(2.70, 1.51)</td>
<td>0.34</td>
<td>(0.73, 1.28)</td>
<td>0.0009</td>
<td>(-0.69, -1.30)</td>
</tr>
<tr>
<td>2</td>
<td>(2.54, 1.21)</td>
<td>0.10</td>
<td>2</td>
<td>(2.44, 1.26)</td>
<td>0.04</td>
<td>(0.18, 0.32)</td>
<td>0.14</td>
<td>(-0.30, -0.25)</td>
</tr>
<tr>
<td>3</td>
<td>(2.25, 1.10)</td>
<td>0.008</td>
<td>1</td>
<td>(2.25, 1.10)</td>
<td>0.008</td>
<td>(0.16, -0.20)</td>
<td>0.32</td>
<td>(-0.16, 0.20)</td>
</tr>
<tr>
<td>4</td>
<td>(2.19, 1.09)</td>
<td>0.0017</td>
<td>2</td>
<td>(2.185, 1.094)</td>
<td>0.0012</td>
<td>(0.002, 0.01)</td>
<td>0.02</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>(2.19, 1.09)</td>
<td>0.0017</td>
<td>2</td>
<td>(2.185, 1.094)</td>
<td>0.0012</td>
<td>(0.002, 0.01)</td>
<td>0.02</td>
<td>—</td>
</tr>
</tbody>
</table>
8.8.8 Theorem

Consider the problem to minimize \( f(x) = c^t x + (1/2)x^t Hx \) subject to \( x \in \mathbb{R}^n \). Suppose that the problem is solved by the conjugate gradient method, starting with \( y_1 \) and letting \( d_1 = -\nabla f(y_1) \). In particular, for \( j = 1, \ldots, n \), let \( \lambda_j \) be an optimal solution to the problem to minimize \( f(y_j + \lambda d_j) \) subject to \( \lambda \geq 0 \). Let \( y_{j+1} = y_j + \lambda_j d_j \), and let \( d_{j+1} = -\nabla f(y_{j+1}) + \alpha_j d_j \), where \( \alpha_j = \frac{\|\nabla f(y_{j+1})\|^2}{\|\nabla f(y_j)\|^2} \). If \( \nabla f(y_j) \neq 0 \) for \( j = 1, \ldots, n \), then the following statements are true:

1. \( d_1, \ldots, d_n \) are H-conjugate.
2. \( d_1, \ldots, d_n \) are descent directions.
3. \( \alpha_j = \frac{\|\nabla f(y_{j+1})\|^2}{\|\nabla f(y_j)\|^2} = \frac{d_j^t H \nabla f(y_{j+1})}{d_j^t H d_j} \) for \( j = 1, \ldots, n \).
First, suppose that Parts 1, 2, and 3 hold true for \( j \). We show that they also hold true for \( j + 1 \). To show that Part 1 holds true for \( j + 1 \), we first demonstrate that \( d_k^t H d_{j+1} = 0 \) for \( k \leq j \). Since \( d_{j+1} = -\nabla f(y_{j+1}) + \alpha_j d_j \), noting the induction hypothesis of Part 3 and letting \( k = j \), we get

\[
d_j^t H d_{j+1} = d_j^t H \left[ -\nabla f(y_{j+1}) + \frac{d_j^t H \nabla f(y_{j+1})}{d_j^t H d_j} d_j \right] = 0. \quad (8.61)
\]

Now let \( k < j \). Since \( d_{j+1} = -\nabla f(y_{j+1}) + \alpha_j d_j \), and since \( d_k^t H d_j = 0 \) by the induction hypothesis of Part 1,

\[
d_{k+1}^t H d_{j+1} = -d_k^t H \nabla f(y_{j+1}). \quad (8.62)
\]

Since \( \nabla f(y_{k+1}) = c + H y_{k+1} \) and \( y_{k+1} = y_k + \lambda_k d_k \), note that

\[
d_{k+1} = -\nabla f(y_{k+1}) + \alpha_k d_k \\
= -[\nabla f(y_k) + \lambda_k H d_k] + \alpha_k d_k \\
= [-d_k + \alpha_k d_{k-1} + \lambda_k H d_k] + \alpha_k d_k.
\]

By the induction hypothesis of Part 2, \( d_k \) is a descent direction and hence, \( \lambda_k > 0 \). Therefore,

\[
d_k^t H = \frac{1}{\lambda_k} [ -d_{k+1}^t + (1 + \alpha_k) d_k - \alpha_{k-1} d_{k-1}]. \quad (8.63)
\]

From (8.62) and (8.63) it follows that

\[
d_{k+1}^t H d_{j+1} = -d_k^t H \nabla f(y_{j+1}) \\
= -\frac{1}{\lambda_k} [-d_{k+1}^t \nabla f(y_{j+1}) + (1 + \alpha_k) d_k^t \nabla f(y_{j+1}) - \alpha_{k-1} d_{k-1} \nabla f(y_{j+1})].
\]

By part 1 of Theorem 8.8.3, and since \( d_1, \ldots, d_j \) are assumed conjugate, \( d_{k+1}^t \nabla f(y_{j+1}) = d_k^t \nabla f(y_{j+1}) = d_{k-1}^t \nabla f(y_{j+1}) = 0 \). Thus, the above equation implies that \( d_k^t H d_{j+1} = 0 \) for \( k < j \). This, together with (8.61), shows that \( d_k^t H d_{j+1} = 0 \) for all \( k \leq j \).

To show that \( d_1, \ldots, d_{j+1} \) are \( H \)-conjugate, it thus suffices to show that they are linearly independent. Suppose that \( \sum_{i=1}^{j+1} \gamma_i d_i = 0 \). Then \( \sum_{i=1}^{j} \gamma_i d_i + \gamma_{j+1} [-\nabla f(y_{j+1}) + \alpha_j d_j] = 0 \). Multiplying by \( \nabla f(y_{j+1})^t \), and noting part 1 of
Theorem 8.8.3, it follows that $\gamma_{j+1} \| \nabla f(y_{j+1}) \|^2 = 0$. Since $\nabla f(y_{j+1}) \neq 0$, $\gamma_{j+1} = 0$. This implies that $\sum_{i=1}^{j} \gamma_i d_i = 0$, and in view of the conjugacy of $d_1, \ldots, d_j$, it follows that $\gamma_1 = \ldots = \gamma_j = 0$. Thus, $d_1, \ldots, d_{j+1}$ are linearly independent and $H$-conjugate, so that Part 1 holds true for $j + 1$.

Now we show that Part 2 holds true for $j + 1$; that is, $d_{j+1}$ is a descent direction. Note that $\nabla f(y_{j+1}) = 0$ by assumption, and that $\nabla f(y_{j+1})' d_j = 0$ by Part 1 of Theorem 8.8.3. Then $\nabla f(y_{j+1})' d_{j+1} = -\| \nabla f(y_{j+1}) \|^2 + \alpha_j \nabla f(y_{j+1})' d_j$

$= -\| \nabla f(y_{j+1}) \|^2 < 0$. By Theorem 4.1.2, $d_{j+1}$ is a descent direction.

Next we show that Part 3 holds true for $j + 1$. By letting $k = j + 1$ in (8.63) and multiplying by $\nabla f(y_{j+2})$, it follows that

$$
\lambda_{j+1} d_{j+1}' H \nabla f(y_{j+2}) = [-d_{j+2}' + (1 + \alpha_{j+1}) d_{j+1}' - \alpha_j d_j'] \nabla f(y_{j+2})
$$

$$
= [\nabla f(y_{j+2})' + d_{j+1}' - \alpha_j d_j'] \nabla f(y_{j+2}).
$$

Since $d_1, \ldots, d_{j+1}$ are $H$-conjugate, then, by Part 1 of Theorem 8.8.3, $d_{j+1}' \nabla f(y_{j+2}) = d_j' \nabla f(y_{j+2}) = 0$. The above equation then implies that

$$
\| \nabla f(y_{j+2}) \|^2 = \lambda_{j+1} d_{j+1}' H \nabla f(y_{j+2}). \tag{8.64}
$$

Multiplying $\nabla f(y_{j+1}) = \nabla f(y_{j+2}) - \lambda_{j+1} H d_{j+1}$ by $\nabla f(y_{j+1})'$, and noting that $d_j' H d_{j+1} = d_{j+1}' \nabla f(y_{j+2}) = d_j' \nabla f(y_{j+2}) = 0$, we get

$$
\| \nabla f(y_{j+1}) \|^2 = \nabla f(y_{j+1})' [\nabla f(y_{j+2}) - \lambda_{j+1} H d_{j+1}]
$$

$$
= (-d_{j+1}' + \alpha_j d_j') [\nabla f(y_{j+2}) - \lambda_{j+1} H d_{j+1}] 
$$

$$
= \lambda_{j+1} d_{j+1}' H d_{j+1}. \tag{8.65}
$$

From (8.64) and (8.65), it is obvious that part 3 holds true for $j + 1$.

We have thus shown that if Parts 1, 2, and 3 hold true for $j$, then they also hold true for $j + 1$. Note that Parts 1 and 2 trivially hold true for $j = 1$. In addition, using an argument similar to that used in proving that Part 3 holds true for $j + 1$, it can easily be demonstrated that it holds true for $j = 1$. This completes the proof.

The reader should note here that when the function $f$ is quadratic and when exact line searches are performed, the choices of $\alpha_j$, given variously by
(8.57), (8.58), and (8.60) all coincide, and thus Theorem 8.8.8 also holds true for the Hestenes and Stiefel (HS) and the Polak and Ribiere (PR) choices of $\alpha_j$. However, for nonquadratic functions, the choice $\alpha_j^{PR}$ appears to be empirically superior to $\alpha_j^{FR}$. This is understandable, since the reduction of (8.58) to (8.60) assumes $f$ to be quadratic. In the same vein, when inexact line searches are performed, the choice $\alpha_j^{HS}$ appears to be preferable. Note that even when $f$ is quadratic, if inexact line searches are performed, the conjugacy relationship holds true only between consecutive directions. We refer the reader to the Notes and References section for a discussion on some alternative three-term recurrence relationships for generating mutually conjugate directions in such a case.

Also note that we have used $d_1 = -IVf(y_1)$ in the foregoing analysis. In lieu of using the identity matrix here, we could have used some general preconditioning matrix $D$, where $D$ is symmetric and positive definite. This would have given $d_1 = -D\nabla f(y_1)$, and (8.56b) would have become $d_{j+1} = -D\nabla f(y_{j+1}) + \alpha_j d_j$, where, for example, in the spirit of (8.57), we have

$$\alpha_j^{HS} = \frac{q_{j} D\nabla f(y_{j+1})}{q_{j}^{\top} d_j}.$$ 

This corresponds, essentially to making a change of variables $y' = D^{-1/2}y$ and using the original conjugate gradient algorithm. Therefore, this motivates the choice of $D$ from the viewpoint of improving the eigenstructure of the problem, as discussed earlier.

For quadratic functions $f$, the conjugate gradient step also has an interesting pattern search interpretation. Consider Figure 8.24 and suppose that the successive points $y_j$, $y_{j+1}$, and $y_{j+2}$ are generated by the conjugate gradient algorithm. Now, suppose that at the point $y_{j+1}$ obtained from $y_j$ by minimizing along $d_j$, we had instead minimized next along the steepest descent direction $-\nabla f(y_{j+1})$ at $y_{j+1}$, leading to the point $y'_{j+1}$. Then it can be shown (see Exercise 8.38) that a pattern search step of minimizing the quadratic function $f$ from $y_j$ along the direction $y'_{j+1} - y_j$ would also have led to the same point $y_{j+2}$. The method, which uses the latter kind of step in general (even for nonquadratic functions), is more popularly known as PARTAN (see Exercise 8.53). Note that the global convergence of PARTAN for general functions is tied into using the negative gradient direction as a spacer step in Theorem 7.3.4 and is independent of any restart conditions, although it is recommended that the method be restarted every $n$ iterations to promote its behavior as a conjugate gradient method.
Memoryless Quasi-Newton Methods

There is an interesting connection between conjugate gradient methods and a simplified variant of the BFGS quasi-Newton method. Suppose that we operate the latter method by updating the inverse Hessian approximation according to $D_{j+1} = D_j + C_j^{BFGS}$, where the correction matrix $C_j^{BFGS}$ is given in (8.48), but assuming that $D_j = I$. Hence, we get

$$D_{j+1} = I + \frac{p_j p_j^T}{p_j^T q_j} \left( 1 + \frac{q_j^T q_j}{p_j^T q_j} \right) - \frac{q_j p_j^T + p_j q_j^T}{p_j^T q_j}. \quad (8.66a)$$

We then move along the direction

$$d_{j+1} = -D_{j+1} \nabla f(y_{j+1}). \quad (8.66b)$$

This is akin to "forgetting" the previous approximation $D_j$ and, instead, updating the identity matrix as might be done at the first iteration of a quasi-Newton method: hence, the name memoryless quasi-Newton method. Observe that the storage requirements are similar to that of conjugate gradient methods and that inexact line searches can be performed as long as $p_j^T q_j = \lambda_j d_j^T [\nabla f(y_{j+1}) - \nabla f(y_j)]$ remains positive and $d_{j+1}$ continues to be a descent direction. Also, note that the loss of positive definiteness of the approximations $D_j$ in the quasi-Newton method is now no longer of concern. In fact, this scheme has proved to be computationally very effective in conjunction with inexact line searches. We refer the reader to the Notes and References section for a discussion on conjugate gradient methods operated with inexact line searches.
Now, suppose that we do employ exact line searches. Then we have
\[ p_j' \nabla f(y_{j+1}) = \lambda_j d_j' \nabla f(y_{j+1}) = 0, \]
so (8.66) gives
\[
\mathbf{d}_{j+1} = -\nabla f(y_{j+1}) + q_j' \nabla f(y_{j+1}) \frac{p_j}{p_j' q_j} \mathbf{p}_j = -\nabla f(y_{j+1}) + \alpha_j^{HS} \mathbf{d}_j
\]
from (8.57). Hence, the BFGS memoryless update scheme is equivalent to the conjugate gradient method of Hestenes and Stiefel (or Polak and Ribiere) when exact line searches are employed. We mention here that although this memoryless update can be performed on any other member of the Broyden family as well (see Exercise 8.34), the equivalence with conjugate gradient methods results only for \( \phi = 1 \) (the BFGS update), as does the observed empirical effectiveness of this scheme (see Exercise 8.40).

Recommendations for Restarting Conjugate Gradient Methods

In several computational experiments using different conjugate gradient techniques, with or without exact line searches, it has been demonstrated time and again that the performance of conjugate gradient methods can be greatly enhanced by employing a proper restart criterion. In particular, a restart procedure suggested by Beale [1970] and augmented by Powell [1977] has proved to be very effective and is invariably implemented, as described below.

Consider the conjugate gradient method summarized formally above in the context of Fletcher and Reeves’s choice of \( \alpha_j \). (Naturally, this strategy applies to any other admissible choice of \( \alpha_j \) as well.) At some inner loop iteration \( j \) of this procedure, having found that \( y_{j+1} = y_j + \lambda_j d_j \) by searching along \( d_j \) from the point \( y_j \), suppose that we decide to reset. (In the previous description of the algorithm, this decision was made whenever \( j = n \).) Let \( \tau = j \) denote this restart iteration. For the next iteration, we find the search direction
\[
\mathbf{d}_{\tau+1} = -\nabla f(y_{\tau+1}) + \alpha_\tau \mathbf{d}_\tau
\]
(8.67)
as usual. Then at Step 3, we replace \( y_1 \) by \( y_{\tau+1} \), let \( x_{k+1} = y_{\tau+1} \), \( d_1 = d_{\tau+1} \), and return to Step 1 to continue with the next set of inner loop iterations. However, instead of computing \( d_{j+1} = -\nabla f(y_{j+1}) + \alpha_j d_j \) for \( j \geq 1 \), we now use
\[
\mathbf{d}_2 = -\nabla f(y_2) + \alpha_1 \mathbf{d}_1
\]
(8.68a)
and
\[
\mathbf{d}_{j+1} = -\nabla f(y_{j+1}) + \alpha_j \mathbf{d}_j + \gamma_j \mathbf{d}_1 \quad \text{for} \ j \geq 2,
\]
where
\[ \gamma_j = \frac{\nabla f(y_{j+1})^T q_1}{d_1^T q_1} \]  

and where \( \alpha_j \) is computed as before, depending on the method being used. Note that (8.68a) employs the usual conjugate gradient scheme, thereby yielding \( d_1 \) and \( d_2 \) as \( H \)-conjugate when \( f \) is quadratic. However, when \( f \) is quadratic with a positive definite Hessian \( H \) and \( d_1 \) is chosen arbitrarily, then when \( j = 2 \), for example, the usual choice of \( \alpha_2 \) would make \( d_3 \) and \( d_2 \) \( H \)-conjugate, but we would need something additional to make \( d_3 \) and \( d_1 \) \( H \)-conjugate. This is accomplished by the extra term \( \gamma_2 d_1 \). Indeed, requiring that \( d_3^T H d_1 = 0 \), where \( d_3 \) is given by the expression in (8.68b), and noting that \( d_2^T H d_1 = 0 \) yields \( \gamma_2 = \nabla f(y_3)^T H d_1 / d_1^T H d_1 = \nabla f(y_3)^T q_1 / d_1^T q_1 \). Proceeding inductively in this manner, the additional term in (8.68b) ensures the \( H \)-conjugacy of all directions generated (see Exercise 8.48).

The foregoing scheme was suggested by Beale with the motivation that whenever a restart is done using \( d_1 = -\nabla f(y_1) \) instead of \( d_1 = d_{r+1} \) as given by (8.67), we lose important second-order information inherent in \( d_r \). Additionally, Powell suggested that after finding \( y_{j+1} \), if any of the following three conditions holds true, then the algorithm should be restarted by putting \( r = j \), computing \( d_{r+1} \) via (8.67), and resetting \( d_1 = d_{r+1} \) and \( y_1 = y_{r+1} \):

1. \( j = n - 1 \).
2. \( \left| \nabla f(y_{j+1})^T \nabla f(y_j) \right| \leq 0.2 \left\| \nabla f(y_{j+1}) \right\|^2 \) for some \( j \geq 1 \).
3. \( -1.2 \left\| \nabla f(y_{j+1}) \right\|^2 \leq d_{j+1}^T \nabla f(y_{j+1}) \leq -0.8 \left\| \nabla f(y_{j+1}) \right\|^2 \) is violated for some \( j \geq 2 \).

Condition 1 is the usual reset criterion by which, after searching along the direction \( d_{r+1} = d_n \), we will have searched along \( n \) conjugate directions for the quadratic case. Condition 2 suggests a reset if a sufficient measure of orthogonality has been lost between \( \nabla f(y_j) \) and \( \nabla f(y_{j+1}) \), motivated by the expanding subspace property illustrated in Figure 8.21. (Computationally, instead of using 0.2 here, any constant in the interval \( [0.1, 0.9] \) appears to give satisfactory performance.) Condition 3 checks for a sufficient descent along the direction \( d_{j+1} \) at the point \( y_{j+1} \), and it also checks for the relative accuracy of the identity \( d_{j+1}^T \nabla f(y_{j+1}) = -\left\| \nabla f(y_{j+1}) \right\|^2 \), which must hold true under exact line searches [whence, using (8.56b), we would have \( d_j^T \nabla f(y_{j+1}) = 0 \)]. For similar ideas when employing inexact line searches, we refer the reader to the Notes and References section.
Convergence of Conjugate Direction Methods

As shown in Theorem 8.8.3, if the function under consideration is quadratic, then any conjugate direction algorithm produces an optimal solution in a finite number of steps. We now discuss the convergence of these methods if the function is not necessarily quadratic.

In Theorem 7.3.4 we showed that a composite algorithm \( A = CB \) converges to a point in the solution set \( \Omega \) if the following properties hold true:

1. \( B \) is closed at points not in \( \Omega \).
2. If \( y \in B(x) \), then \( f(y) < f(x) \) for \( x \notin \Omega \).
3. If \( z \in C(y) \), then \( f(z) \leq f(y) \).
4. The set \( \Lambda = \{ x : f(x) \leq f(x_1) \} \) is compact, where \( x_1 \) is the starting solution.

For the conjugate direction (quasi-Newton or conjugate gradient) algorithms discussed in this chapter, the map \( B \) is of the following form. Given \( x \), then \( y \in B(x) \) means that \( y \) is obtained by minimizing \( f \) starting from \( x \) along the direction \( d = -DVf(x) \), where \( D \) is a specified positive definite matrix. In particular, for the conjugate gradient methods, \( D = I \), and for the quasi-Newton methods, \( D \) is an arbitrary positive definite matrix. Furthermore, starting from the point obtained by applying the map \( B \), the map \( C \) is defined by minimizing the function \( f \) along the directions specified by the particular algorithms. Thus, the map \( C \) satisfies Property 3.

Letting \( \Omega = \{ x : \nabla f(x) = 0 \} \), we now show that the map \( B \) satisfies Properties 1 and 2. Let \( x \in \Omega \) and let \( x_k \to x \). Furthermore, let \( y_k \in B(x_k) \) and let \( y_k \to y \). We need to show that \( y \in B(x) \). By the definition of \( y_k \), we have \( y_k = x_k - \lambda_k D\nabla f(x_k) \) for \( \lambda_k \geq 0 \) such that

\[
f(y_k) \leq f(x_k - \lambda D\nabla f(x_k)) \quad \text{for all } \lambda \geq 0.
\] (8.69)

Since \( \nabla f(x) \neq 0 \), then \( \lambda_k \) converges to \( \lambda = \| y - x \| / \| D\nabla f(x) \| \geq 0 \). Therefore, \( y = x - \lambda D\nabla f(x) \). Taking the limit as \( k \to \infty \) in (8.69), \( f(y) = f[x - \lambda D\nabla f(x)] \) for all \( \lambda \geq 0 \), so that \( y \) is indeed obtained by minimizing \( f \) starting from \( x \) in the direction \( -D\nabla f(x) \). Thus, \( y \in B(x) \), and \( B \) is closed. Also, Part 2 holds true by noting that \( -\nabla f(x)^T D\nabla f(x) < 0 \), so that \( -D\nabla f(x) \) is a descent direction. Assuming that the set defined in Part 4 is compact, it follows that the conjugate direction algorithms discussed in this section converge to a point with zero gradient.

The role played by the map \( B \) described above is akin to that of a spacer step, as discussed in connection with Theorem 7.3.4. For algorithms that are designed empirically and that may not enjoy theoretical convergence, this can be alleviated by inserting such a spacer step involving a periodic minimization
along the negative gradient direction, for example, hence, achieving theoretical convergence.

We now turn our attention to addressing the rate of convergence or local convergence characteristics of the algorithms discussed in this section.

**Convergence Rate Characteristics for Conjugate Gradient Methods**

Consider the quadratic function \( f(x) = c^T x + \frac{1}{2} x^T H x \), where \( H \) is an \( n \times n \) symmetric, positive definite matrix. Suppose that the eigenvalues of \( H \) are grouped into two sets, of which one set is composed of some \( m \) relatively large and perhaps dispersed values, and the other set is a cluster of some \( n - m \) relatively smaller eigenvalues. (Such a structure arises, for example, with the use of quadratic penalty functions for linearly constrained quadratic programs, as discussed in Chapter 9.) Let us assume that \( (m + 1) \leq n \), and let \( \alpha \) denote the ratio of the largest to the smallest eigenvalue in the latter cluster. Now, we know that a standard application of the conjugate gradient method will result in a finite convergence to the optimum in \( n \), or fewer, steps. However, suppose that we operate the conjugate gradient algorithm by restarting with the steepest descent direction every \( m + 1 \) line searches or steps. Such a procedure is called a *partial conjugate gradient method*.

Starting with a solution \( x_1 \), let \( \{x_k\} \) be the sequence thus generated, where for each \( k \geq 1 \), \( x_{k+1} \) is obtained after applying \( m + 1 \) conjugate gradient steps upon restarting with \( x_k \) as above. Let us refer to this as an \( (m + 1) \)-step process. As in Equation (8.17), let us define an error function \( e(x) = \frac{1}{2}(x - x^*)^T H (x - x^*) \), which differs from \( f(x) \) by a constant, and which is zero if and only if \( x = x^* \). Then it can be shown (see the Notes and References section) that

\[
e(x_{k+1}) \leq \frac{(\alpha - 1)^2}{(\alpha + 1)^2} e(x_k).
\]  

(8.70)

Hence, this establishes a linear rate of convergence for the above process as in the special case of the steepest descent method for which \( m = 0 \) [see Equation (8.18)]. However, the ratio \( \alpha \) that governs the convergence rate is now independent of the \( m \) largest eigenvalues. Thus, the effect of the \( m \) largest eigenvalues is eliminated, but at the expense of an \( (m + 1) \)-step process versus the single-step process of the steepest descent method.

Next, consider the general nonquadratic case to which the usual \( n \)-step conjugate gradient process is applied. Intuitively, since the conjugate gradient method accomplishes in \( n \) steps what Newton's method does in a single step, by the local quadratic convergence rate of Newton's method, we might similarly expect that the \( n \)-step conjugate gradient process also converges quadratically; that is, \( \|x_{k+1} - x^*\| \leq \beta \|x_k - x^*\|^2 \) for some \( \beta > 0 \). Indeed, it can be shown (see
the Notes and References section) that if the sequence \( \{x_k\} \to x^* \), the function under consideration is twice continuously differentiable in some neighborhood of \( x^* \), and the Hessian matrix at \( x^* \) is positive definite, the \( n \)-step process converges superlinearly to \( x^* \). Moreover, if the Hessian matrix satisfies an appropriate Lipschitz condition in some neighborhood of \( x^* \), then the rate of superlinear convergence is \( n \)-step quadratic. Again, caution must be exercised in interpreting these results in comparison with, say, the linear convergence rate of steepest descent methods. That is, these are \( n \)-step asymptotic results, whereas the steepest descent method is a single-step procedure. Also, given that these methods are usually applied when \( n \) is relatively large, it is seldom practical to perform more than \( 5n \) iterations, or five \( n \)-step iterations. Fortunately, empirical results seem to indicate that this does not pose a problem because reasonable convergence is typically obtained within \( 2n \) iterations.

**Convergence Rate Characteristics for Quasi-Newton Methods**

The Broyden class of quasi-Newton methods can also be operated as partial quasi-Newton methods by restarting every \( m + 1 \) iterations with, say, the steepest descent direction. For the quadratic case, the local convergence properties of such a scheme resembles that for conjugate gradient methods as discussed above. Also, for nonquadratic cases, the \( n \)-step quasi-Newton algorithm has a local superlinear convergence rate behavior similar to that of the conjugate gradient method. Intuitively, this is because of the identical effect that the \( n \)-step process of either method has on quadratic functions. Again, the usual caution must be adopted in interpreting the value of an \( n \)-step superlinear convergence behavior. Additionally, we draw the reader's attention to Exercise 8.52 and to the section on scaling quasi-Newton methods, where we discuss the possible ill-conditioning effects resulting from the sequential transformation of the eigenvalues of \( D_{y+1}H \) to unity for the quadratic case.

Quasi-Newton methods are also sometimes operated as a continuing updating process, without resets. Although the global convergence of such a scheme requires rather stringent conditions, the local convergence rate behavior is often asymptotically superlinear. For example, for the BFGS update scheme, which has been seen to exhibit a relatively superior empirical performance, as mentioned previously, the following result holds true (see the Notes and References section). Let \( y^* \) be such that the Hessian \( H(y^*) \) is positive definite and that there exists an \( \varepsilon \)-neighborhood \( N_\varepsilon(y^*) \) of \( y^* \) such that the Lipschitz condition \( \|H(y) - H(y^*)\| \leq L\|y - y^*\| \) holds true for \( y \in N_\varepsilon(y^*) \), where \( L \) is a positive constant. Then, if a sequence \( \{y_k\} \) generated by a continually updated quasi-Newton process with a fixed step size of unity converges to such a \( y^* \), the asymptotic rate of convergence is superlinear. Similar superlinear convergence rate results are available for the DFP algorithm, with both exact line searches
and unit step size choices under appropriate conditions. We refer the reader to the Notes and References section for further reading on this subject.

8.9 Subgradient Optimization

Consider Problem P, defined as

\[ P: \text{Minimize} \ \{f(x) : x \in X\}, \tag{8.71} \]

where \( f: \mathbb{R}^n \to \mathbb{R} \) is a convex but not necessarily differentiable function and where \( X \) is a nonempty, closed, convex subset of \( \mathbb{R}^n \). We assume that an optimal solution exists, as it would be, for example, if \( X \) is bounded or if \( f(x) \to \infty \) whenever \( \|x\| \to \infty \).

For such a Problem P, we now describe a subgradient optimization algorithm that can be viewed as a direct generalization of the steepest descent algorithm in which the negative gradient direction is substituted by a negative subgradient-based direction. However, the latter direction need not necessarily be a descent direction, although, as we shall see, it does result in the new iterate approaching closer to an optimal solution for a sufficiently small step size. For this reason we do not perform a line search along the negative subgradient direction, but rather, we prescribe a step size at each iteration that guarantees that the sequence generated will eventually converge to an optimal solution. Also, given an iterate \( x_k \in X \) and adopting a step size \( \lambda_k \) along the direction \( d_k = -\xi_k / \|\xi_k\| \), where \( \xi_k \) belongs to the subdifferential \( \partial f(x_k) \) of \( f \) at \( x_k \) (\( \xi_k \neq 0 \), say), the resulting point \( \bar{x}_{k+1} = x_k + \lambda_k d_k \) need not belong to \( X \). Consequently, the new iterate \( x_{k+1} \) is obtained by projecting \( \bar{x}_{k+1} \) onto \( X \), that is, finding the (unique) closest point in \( X \) to \( \bar{x}_{k+1} \). We denote this operation as \( x_{k+1} = P_X(\bar{x}_{k+1}) \), where

\[ P_X(\bar{x}) = \text{argmin}\{\|x - \bar{x}\| : x \in X\}. \tag{8.72} \]

The foregoing projection operation should be easy to perform if the method is to be computationally viable. For example, in the context of Lagrangian duality (Chapter 6), wherein subgradient methods and their variants are most frequently used, the set \( X \) might simply represent nonnegativity restrictions \( x \geq 0 \) on the variables. In this case, we easily obtain \( (x_{k+1})_i = \max\{0, (\bar{x}_{k+1})_i\} \) for each component \( i = 1, \ldots, n \) in (8.72). In other contexts, the set \( X = \{x : \ell_i \leq x_i \leq u_i, \ i = 1, \ldots, n\} \) might represent simple finite lower and upper bounds on the variables. In this case, it is again easy to verify that

\[
(x_{k+1})_i = \begin{cases} 
(\bar{x}_{k+1})_i & \text{if } \ell_i \leq (\bar{x}_{k+1})_i \leq u_i \\
\ell_i & \text{if } (\bar{x}_{k+1})_i < \ell_i \\
u_i & \text{if } (\bar{x}_{k+1})_i > u_i 
\end{cases} \text{ for } i = 1, \ldots, n. \tag{8.73}
\]
Also, when an additional knapsack constraint \( a'x = \beta \) is introduced to define \( X = \{ x : a'x = \beta, \ \ell \leq x \leq u \} \), then, again, \( P_X(\bar{x}) \) is relatively easy to obtain (see Exercise 8.60).

Summary of a (Rudimentary) Subgradient Algorithm

**Initialization Step** Select a starting solution \( x_1 \in X \), let the current upper bound on the optimal objective value be \( UB_1 = f(x_1) \), and let the current incumbent solution be \( x^* = x_1 \). Put \( k = 1 \), and go to the Main Step.

**Main Step** Given \( x_k \), find a subgradient \( \xi_k \in \partial f(x_k) \) of \( f \) at \( x_k \). If \( \xi_k = 0 \), then stop; \( x_k \) (or \( x^* \)) solves Problem P. Otherwise, let \( d_k = -\xi_k / \| \xi_k \| \), select a step size \( \lambda_k > 0 \), and compute \( x_{k+1} = P_X(\bar{x}_{k+1}) \), where \( \bar{x}_{k+1} = x_k + \lambda_k d_k \). If \( f(x_{k+1}) < UB_k \), put \( UB_{k+1} = f(x_{k+1}) \) and \( x^* = x_{k+1} \). Otherwise, let \( UB_{k+1} = UB_k \). Increment \( k \) by 1 and repeat the Main Step.

Note that the stopping criterion \( \xi_k = 0 \) may never be realized, even if there exists an interior point optimum and we do find a solution \( x_k \) for which 0 \( \in \partial f(x_k) \), because the algorithm arbitrarily selects the subgradient \( \xi_k \). Hence, a practical stopping criterion based on a maximum limit on the number of iterations performed is used almost invariably. Note also that we can terminate the procedure whenever \( x_{k+1} = x_k \) for any iteration. Alternatively, if the optimal objective value \( f^* \) is known, as in the problem of finding a feasible solution by minimizing the sum of (absolute) constraint violations, an \( \varepsilon \) stopping criterion \( UB_k \leq f^* + \varepsilon \) may be used for some tolerance \( \varepsilon > 0 \). (See the Notes and References section for a primal–dual scheme employing a termination criterion based on the duality gap.)

**8.9.1 Example**

Consider the following Problem P:

\[
\text{Minimize } \{ f(x, y) : -1 \leq x \leq 1, -1 \leq y \leq 1 \}
\]

where \( f(x, y) = \max\{-x, x + y, x - 2y\} \).

By considering \( f(x, y) \leq c \), where \( c \) is a constant, and examining the region bounded by \( -x \leq c, x + y \leq c, \) and \( x - 2y \leq c \), we can plot the contours of \( f \) as shown in Figure 8.25. Note that the points of nondifferentiability are of the type \((t, 0), (-t, 2t), \) and \((-t, -t)\) for \( t \geq 0 \). Also, the optimal solution is \((x, y) = (0, 0)\), at which all three linear functions defining \( f \) tie in value. Hence, although \((0, 0)' \in \)
Unconstrained Optimization

Now consider the point \((x, y) = (1, 0)\). We have \(f(1, 0) = 1\), as determined by the linear functions \(x + y\) and \(x - 2y\). (See Figure 8.25.) Hence, \(\xi = (1, 1)' \in \partial f(1, 0)\). Consider the direction \(-\xi = (-1, -1)'\). Note that this is not a descent direction. However, as we begin to move along this direction, we do approach closer to the optimal solution \((0, 0)'\). Figure 8.25 shows the ideal step that we could take along the direction \(d = -\xi\) to arrive closest to the optimal solution. However, suppose that we take a step length \(\lambda = 2\) along \(-\xi\). This brings us to the point \((1, 0) - 2(1, 1) = (-1, -2)\). The projection \(P_X(-1, -2)\) of \((-1, -2)\) onto \(X\) is obtained via (8.73) as \((-1, -1)\). This constitutes one iteration of the foregoing algorithm.

The following result prescribes a step-size selection scheme that will guarantee convergence to an optimum.

### 8.9.2 Theorem

Let Problem \(P\) be as defined in (8.71) and assume that an optimum exists. Consider the foregoing subgradient optimization algorithm to solve Problem \(P\), and suppose that the prescribed nonnegative step size sequence \(\{\lambda_k\}\) satisfies the conditions \(\lambda_k \to 0^+\) and \(\sum_{k=0}^{\infty} \lambda_k = \infty\). Then, either the algorithm terminates finitely with an optimal solution, or else an infinite sequence is generated such that

![Figure 8.25 Contours of \(f\), in Example 8.9.1.](image)

\{UB_k\} \rightarrow f^* = \min\{f(x) : x \in X\}.

**Proof**

The case of finite termination follows from Theorem 3.4.3. Hence, suppose that an infinite sequence \(\{x_k\}\) is generated along with the accompanying sequence of upper bounds \(\{UB_k\}\). Since \(\{UB_k\}\) is monotone nonincreasing, it has a limit point \(\bar{f}\). We show that this limit \(\bar{f}\) equals \(f^*\) by exhibiting that for any given value \(\alpha > f^*\), the sequence \(\{x_k\}\) enters the level set \(S_\alpha = \{x : f(x) \leq \alpha\}\). Hence, we cannot have \(\bar{f} > f^*\), or else we would obtain a contradiction by taking \(\alpha \in (f^*, \bar{f})\), so we must then have \(\bar{f} = f^*\).

Toward this end, consider any \(\hat{x} \in X\) such that \(f(\hat{x}) < \alpha\). (For example, we can take \(\hat{x}\) as an optimal solution to Problem P.) Since \(\hat{x} \in \text{int } S_\alpha\) because \(f\) is continuous, there exists a \(\rho > 0\) such that \(\|x - \hat{x}\| \leq \rho\) implies that \(x \in S_\alpha\). In particular, \(x_{Bk} = \hat{x} + \rho \xi_k / \|\xi_k\|\) lies on the boundary of the ball centered at \(\hat{x}\) with radius \(\rho\) and hence lies in \(S_\alpha\) for all \(k\). But by the convexity of \(f\), we have \(f(x_{Bk}) \geq f(x_k) + (x_{Bk} - x_k)'\xi_k\) for all \(k\). Hence, on the contrary, if \(\{x_k\}\) never enters \(S_\alpha\), that is, \(f(x_k) > \alpha\) for all \(k\), we shall have \((x_{Bk} - x_k)'\xi_k \leq f(x_{Bk}) - f(x_k) < 0\). Substituting for \(x_{Bk}\), this gives \((\hat{x} - x_k)'\xi_k < -\rho\|\xi_k\|\). Hence, using \(d_k = -\xi_k / \|\xi_k\|\), we get

\[(x_k - \hat{x})'d_k < -\rho\quad \text{for all } k.\quad (8.74)\]

Now we have

\[
\|\bar{x}_{k+1} - \hat{x}\|^2 = \|\bar{x}_{k+1} - x_{k+1} + x_{k+1} - \hat{x}\|^2
\]
\[
= \|x_{k+1} - \hat{x}\|^2 + \|\bar{x}_{k+1} - x_{k+1}\|^2 + 2(\bar{x}_{k+1} - x_{k+1})'(x_{k+1} - \hat{x}).
\]

Hence, by Theorem 2.4.1,

\[
\|\bar{x}_{k+1} - \hat{x}\|^2 = \|x_{k+1} - \hat{x}\|^2 - \|\bar{x}_{k+1} - x_{k+1}\|^2 - 2(\bar{x}_{k+1} - x_{k+1})'(x_{k+1} - \hat{x})
\]
\[
\leq \|\bar{x}_{k+1} - \hat{x}\|^2.
\]

Hence, we get

\[
\|x_{k+1} - \hat{x}\|^2 \leq \|\bar{x}_{k+1} - \hat{x}\|^2 = \|x_k + \lambda_k d_k - \hat{x}\|^2
\]
\[
= \|x_k - \hat{x}\|^2 + \lambda_k^2 + 2\lambda_k d_k'(x_k - \hat{x}).
\]
Using (8.74), this gives

\[ \|x_{k+1} - \hat{x}\|^2 \leq \|x_k - \hat{x}\|^2 + \lambda_k (\lambda_k - 2\rho). \]

Since \( \lambda_k \to 0^+ \), there exists a \( K \) such that for \( k \geq K \), we have \( \lambda_k \leq \rho \). Hence,

\[ \|x_{k+1} - \hat{x}\|^2 \leq \|x_k - \hat{x}\|^2 - \rho \lambda_k \quad \text{for all } k \geq K. \quad (8.75) \]

Summing the inequalities (8.75) written for \( k = K, K + 1, \ldots, K + r \), say, we get

\[ \rho \sum_{k=K}^{K+r} \lambda_k \leq \|x_K - \hat{x}\|^2 - \|x_{K+r+1} - \hat{x}\|^2 \leq \|x_K - \hat{x}\|^2 \quad \text{for all } r \geq 0. \]

Since the sum on the left-hand side diverges to infinity as \( r \to \infty \), this leads to a contradiction, and the proof is complete.

Note that the proof of the theorem can easily be modified to show that for each \( \alpha > f^* \), the sequence \( \{x_k\} \) enters \( S_\alpha \) infinitely often or else, for some \( K' \), we would have \( f(x_k) > \alpha \) for all \( k \geq K' \), leading to the same contradiction. Hence, whenever \( x_{k+1} = x_k \) in the foregoing algorithm, \( x_k \) must be an optimal solution.

Furthermore, the above algorithm and proof can be extended readily to solve the problem of minimizing \( f(x) \) subject to \( x \in X \cap Q \), where \( f \) and \( X \) are as above and where \( Q = \{x : g_i(x) \leq 0, i = 1, \ldots, m\} \). Here, we assume that each \( g_i, i = 1, \ldots, m \), is convex and that \( X \cap \text{int}(Q) \neq \emptyset \), so that for each \( \alpha > f^* \), by defining \( S_\alpha = \{x \in Q : f(x) \leq \alpha\} \), we have a point \( \hat{x} \in X \cap \text{int}(S_\alpha) \). Now, in the algorithm, if we let \( \xi_k \) be a subgradient of \( f \) whenever \( x_k \in Q \), and if we let \( \xi_k \) be a subgradient of the most violated constraint in \( Q \) if \( x_k \not\in Q \) (noting that \( x_k \) always lies in \( X \) by virtue of the projection operation), we shall again have (8.74) holding true, and the remainder of the convergence proof would then follow as before.

**Choice of Step Sizes**

Theorem 8.2.9 guarantees that as long as the step sizes \( \lambda_k, \forall k \), satisfy the conditions stated, convergence to an optimal solution will be obtained. Although this is true theoretically, it is unfortunately far from what is realized in practice. For example, choosing \( \lambda_k = 1/k \) according to the divergent harmonic series \( \sum_{k=1}^{\infty} (1/k) = \infty \), the algorithm can easily stall and be remote from optimality after thousands of iterations. A careful fine tuning of the choice of step sizes is usually required to obtain a satisfactory algorithmic performance.
To gain some insight into the choice of step sizes, let $x_k$ be a nonoptimal iterate with $\xi_k \in \partial f(x_k)$ and denote by $x^*$ an optimal solution to Problem (8.71) having objective value $f^* = f(x^*)$. By the convexity of $f$, we have $f(x^*) \geq f(x_k) + (x^* - x_k)^T \xi_k$, or $(x^* - x_k)^T (-\xi_k) \geq f(x_k) - f^* > 0$. Hence, as observed in Example 8.9.1 (see Figure 8.25), although the direction $d_k = -\xi_k / \|\xi_k\|$ need not be an improving direction, it does lead to points that are closer in Euclidean norm to $x^*$ than was $x_k$. In fact, this is the feature that drives the convergence of the algorithm and ensures an eventual improvement in objective function value.

Now, as in Figure 8.25, an ideal step size to adopt might be that which brings us closest to $x^*$. This step size $\lambda_k^*$ can be found by requiring that the vector $(x_k + \lambda_k^* d_k) - x^*$ is orthogonal to $d_k$, or that $d_k^T [x_k + \lambda_k^* d_k - x^*] = 0$. This gives

$$\lambda_k^* = (x^* - x_k)^T d_k = \frac{(x_k - x^*)^T \xi_k}{\|\xi_k\|}.$$  

Of course, the problem with trying to implement this step size $\lambda_k^*$ is that $x^*$ is unknown. However, by the convexity of $f$, we have $f^* = f(x^*) \geq f(x_k) + (x^* - x_k)^T \xi_k$. Hence, from (8.76), we have that $\lambda_k^* \geq \frac{|f(x_k) - f^*|}{\|\xi_k\|}$. Since $f^*$ is also usually unknown, we can recommend using an underestimate, $\tilde{f}$, in lieu of $f^*$, noting that the foregoing relationship is a “greater than or equal to” type of inequality. This leads to a choice of step size

$$\lambda_k = \frac{\beta_k [f(x_k) - \tilde{f}]}{\|\xi_k\|},$$  

where $\beta_k > 0$. In fact, by selecting $\varepsilon_1 < \beta_k \leq 2 - \varepsilon_2$ for all $k$ for some positive $\varepsilon_1$ and $\varepsilon_2$, and using $f^*$ itself instead of $\tilde{f}$ in (8.77), it can be shown that the generated sequence $\{x_k\}$ converges to an optimum $x^*$. (A linear or geometric convergence rate can be exhibited under some additional assumptions on $f$.)

A practical way of employing (8.77) that has been found empirically to be computationally attractive is as follows (this is called a block halving scheme). First, designate an upper limit $N$ on the number of iterations to be performed. Next, select some $\tilde{r} < N$ and divide the potential sequence of iterations $1, \ldots, N$ into $T = \lceil N / \tilde{r} \rceil$ blocks, with the first $T - 1$ blocks having $\tilde{r}$ iterations, and the final block having the remaining ($\leq \tilde{r}$) iterations. Also, for
each block \( t \), select a parameter value \( \beta(t) \), for \( t = 1, \ldots, T \). [Typical values are \( N = 200, \bar{F} = 75, \beta(1) = 0.75, \beta(2) = 0.5, \) and \( \beta(3) = 0.25, \) with \( T = 3 \).] Now, within each block \( t \), compute the first step length using (8.77), with \( \beta_k \) equal to the corresponding \( \beta(t) \) value. However, for the remaining iterations within the block, the step length is kept the same as for the initial iteration in that block, except that each time the objective function value fails to improve over some \( \bar{V} \) (= 10, say) consecutive iterations, the step length is successively halved. [Alternatively, (8.77) can be used to compute the step length for each iteration, with \( \beta_k \) starting at \( \beta(t) \) for block \( t \), and with this \( \beta \) parameter being halved whenever the method experiences \( \bar{V} \) consecutive failures as before.] Additionally, at the beginning of a new block, and also whenever the method experiences \( \bar{V} \) consecutive failures, the process is reset to the incumbent solution before the modified step length is used. Although some fine tuning of the foregoing parameter values might be required, depending on the class of problems being solved, the prescribed values work well on reasonably well-scaled problems (see the Notes and References section for empirical evidence using such a scheme).

**Subgradient Deflection, Cutting Plane, and Variable Target Value Methods**

It has frequently been observed that the difficulty associated with subgradient methods is that as the iterates progress, the angle between the subgradient-based direction \( d_k \) and the direction \( x^* - x_k \) toward optimality, although acute, tends to approach 90°. As a result, the step size needs to shrink considerably before a descent is realized, and this, in turn, causes the procedure to stall. Hence, it becomes almost imperative to adopt some suitable deflection or rotation scheme to accelerate the convergence behavior.

Toward this end, in the spirit of conjugate gradient methods, we could adopt a direction of search as \( d_1 = -\xi_1 \) and \( d_k = -\xi_k + \phi_k d_{k-1}^2 \), where \( d_{k-1}^2 = x_k - x_{k-1} \) and \( \phi_k \) is an appropriate parameter. (These directions can be normalized and then used in conjunction with the same block-halving step size strategy described above.) Various strategies prompted by theoretical convergence and/or practical efficiency can be designed by choosing \( \phi_k \) appropriately (see the Notes and References section). A simple choice that works reasonably well in practice is the average direction strategy, for which \( \phi_k = \| \xi_k \| / \| d_{k-1}^2 \| \), so that \( d_k \) bisects the angle between \( -\xi_k \) and \( d_{k-1}^2 \).

Another viable strategy is to imitate quasi-Newton procedures by using \( d_k = -D_k \xi_k \), where \( D_k \) is a suitable, symmetric, positive definite matrix. This leads to the class of space dilation methods (see the Notes and References section). Alternatively, we could generate a search direction by finding the minimum norm subgradient as motivated by Theorem 6.3.11, but based on an
approximation to the subdifferential at \( x_k \) and not the actual subdifferential as in the theorem. The class of bundle methods (see the Notes and References section) are designed to iteratively refine such an approximation to the subdifferential until the least norm element yields a descent direction. Note that this desirable strict descent property comes at the expense of having to solve quadratic optimization subproblems, which detract from the simplicity of the foregoing types of subgradient methods.

Thus far, the basic algorithm scheme that we have adopted involves first finding a direction of motion \( d_k \) at a given iterate \( x_k \), followed by computing a prescribed step size \( \lambda_k \) in order to determine the next iterate according to

\[
x_{k+1} = P_X(\bar{x}_{k+1}), \quad \text{where} \quad \bar{x}_{k+1} = x_k + \lambda_k d_k.
\]

There exists an alternative approach in which \( x_{k+1} \) is determined directly via a projection of \( x_k \) onto the polyhedron defined by one or more cutting planes, thereby effectively yielding the direction and step size simultaneously. To motivate this strategy, let us first consider the case of a single cutting plane. Note that by the assumed convexity of \( f \) we have

\[
f(x) \leq f(x_k) + (x - x_k)^T \xi_k,
\]

where \( \xi_k \in \partial f(x_k) \). Let \( f^* \) denote the optimal objective value, and assume for now that \( f(x_k) > f^* \), so that \( \xi_k \) is nonzero. Consider the Polyak–Kelly cutting plane generated from the foregoing convexity-based inequality by imposing the desired restriction that \( f(x) \leq f^* \) as given by

\[
(x - x_k)^T \xi_k \leq f^* - f(x_k). \quad (8.78)
\]

Observe that the current iterate \( x_k \) violates this inequality since \( f(x_k) > f^* \), and hence, (8.78) constitutes a cutting plane that deletes \( x_k \). If we were to project the point \( x_k \) onto this cutting plane, we would effectively move from \( x_k \) a step length of \( \lambda \), say, along the negative normalized gradient \( d_k = -\xi_k / \|\xi_k\| \), such that \( x_k + \lambda d_k \) satisfies (8.78) as an equality. This yields the projected solution

\[
\bar{x}_{k+1} = x_k + \lambda d_k, \quad \text{where} \quad d_k = -\xi_k / \|\xi_k\| \quad \text{and} \quad \lambda = \frac{f(x_k) - f^*}{\|\xi_k\|}. \quad (8.79)
\]

Observe that the effective step length \( \lambda \) in (8.79) is of the type given in (8.77), with \( f^* \) itself being used in lieu of the underestimate \( \overline{f} \), and with \( \beta_k = 1 \). This affords another interpretation for the step size (8.77).

Following this concept, we now concurrently examine a pair of Polyak–Kelly cutting planes based on the current and previous iterates \( x_k \) and \( x_{k-1} \),
respectively. These cuts are predicated on some underestimate \( \bar{f}_k \) at the present iteration \( k \) that is less than the currently best known objective value. Imitating (8.78), this yields the set

\[
G_k = \{ x : (x-x_j)'\xi_j \leq \bar{f}_k - f(x_j) \text{ for } j = k-1 \text{ and } k \}. \tag{8.80}
\]

We then compose the next iterate via a projection onto the polyhedron \( G_k \) according to

\[
x_{k+1} = P_X(\bar{x}_{k+1}), \quad \text{where } \bar{x}_{k+1} = P_{G_k}(x_k). \tag{8.81}
\]

Because of its simple two-constraint structure, the projection \( P_{G_k}(\cdot) \) is relatively easy to compute in closed-form by examining the KKT conditions for the underlying projection problem (see Exercise 8.58). This process of determining the direction and step size simultaneously has been found to be computationally very effective, and can be proven to converge to an optimum under an appropriate prescription of \( \bar{f}_k \), \( \forall k \) (see below as well as the Notes and References section).

We conclude this section by providing an important relevant comment on selecting a suitable underestimating value \( \bar{f}_k \), \( \forall k \), that could be used in place of \( f \) within (8.77), or in the algorithmic process described by (8.80) and (8.81). Note that, in general, we typically do not have any prior information on such a suitable lower bound on the problem. It is of interest, therefore, to design algorithms that would prescribe an automatic scheme for generating and iteratively manipulating an estimate \( \bar{f}_k \) for \( f^* \), \( \forall k \), that in concert with prescribed direction-finding and step-size schemes would ensure that \( \{ \bar{f}_k \} \to f^* \) and \( \{ x_k \} \to x^* \) (over some convergent subsequence) as \( k \to \infty \). There exists a class of algorithms called \textit{variable target value methods} that possesses this feature. Note that the estimate \( \bar{f}_k \) at any iteration \( k \) in these procedures might not be a true underestimate for \( f^* \). Rather, \( \bar{f}_k \) merely serves as a current \textit{target value} to be achieved, which happens to be less than the objective function value best known at present. The idea, then, is to decrease or increase \( \bar{f}_k \) suitably, depending on whether or not a defined sufficient extent of progress is made by the algorithm, in a manner that finally induces convergence to an optimal solution. Approaches of this type have been designed to yield both theoretically convergent and practically effective schemes under various deflected subgradient and step-size schemes, including cutting plane projection methods as described above. We refer the reader to the Notes and References section for a further study on this subject.
Exercises

[8.1] Find the minimum of $6e^{-2\lambda} + 2\lambda^2$ by each of the following procedures:
   a. Golden section method.
   b. Dichotomous search method.
   c. Newton's method.
   d. Bisection search method.

[8.2] For the uniform search method, the dichotomous search method, the golden section method, and the Fibonacci search method, compute the number of functional evaluations required for $\alpha = 0.1, 0.01, 0.001,$ and $0.0001$, where $\alpha$ is the ratio of the final interval of uncertainty to the length of the initial interval of uncertainty.

[8.3] Consider the function $f$ defined by $f(x) = (x_1 + x_2^3)^2 + 2(x_1 - x_2 - 4)^4$. Given a point $x_1$ and a nonzero direction vector $d$, let $\theta(\lambda) = f(x_1 + \lambda d)$.
   a. Obtain an explicit expression for $\theta(\lambda)$.
   b. For $x_1 = (0, 0)^T$ and $d = (1, 1)^T$, using the Fibonacci method, find the value of $\lambda$ that solves the problem to minimize $\theta(\lambda)$ subject to $\lambda \in R$.
   c. For $x_1 = (5, 4)^T$ and $d = (-2, 1)^T$, using the golden section method, find the value of $\lambda$ that solves the problem to minimize $\theta(\lambda)$ subject to $\lambda \in R$.
   d. Repeat parts b and c using the interval bisection method.

[8.4] Show that the method of Fibonacci approaches the golden section method as the number of functional evaluations $n$ approaches $\infty$.

[8.5] Consider the problem to minimize $f(x + \lambda d)$ subject to $\lambda \in R$. Show that a necessary condition for a minimum at $\hat{\lambda}$ is that $d^T \nabla f(y) = 0$, where $y = x + \hat{\lambda} d$. Under what assumptions is this condition sufficient for optimality?

[8.6] Suppose that $\theta$ is differentiable, and let $|\theta'| \leq \alpha$. Furthermore, suppose that the uniform search method is used to minimize $\theta$. Let $\hat{\lambda}$ be a grid point such that $\theta(\lambda) - \theta(\hat{\lambda}) \geq \epsilon > 0$ for each grid point $\lambda \neq \hat{\lambda}$. If the grid length $\delta$ is such that $a\delta \leq \epsilon$, show, without assuming strict quasiconvexity, that no point outside the interval $[\hat{\lambda} - \delta, \hat{\lambda} + \delta]$ could provide a functional value of less than $\theta(\hat{\lambda})$.

[8.7] Consider the problem to minimize $f(x + \lambda d)$ subject to $x + \lambda d \in S$ and $\lambda \geq 0$, where $S$ is a compact convex set and $f$ is a convex function. Furthermore, suppose that $d$ is an improving direction. Show that an optimal solution $\bar{x}$ is
given by \( \bar{\lambda} = \min\{\lambda_1, \lambda_2\} \), where \( \lambda_1 \) satisfies \( d' \nabla f(x + \lambda_1 d) = 0 \), and \( \lambda_2 = \max\{\lambda : x + \lambda d \in S\} \).

[8.8] Define the percentage test line search map that determines the step length \( \lambda \) to within 100\% \( p \), \( 0 \leq p \leq 1 \), of the ideal step \( \lambda^* \) according to \( M(x, d) = \{y : y = x + \lambda d, \text{ where } 0 \leq \lambda < \infty, \text{ and } |\lambda - \lambda^*| \leq p \lambda^*\} \), where defining \( \Theta(\lambda) = f(x + \lambda d) \), we have \( \Theta'(\lambda^*) = 0 \). Show that if \( d \neq 0 \) and \( \Theta \) is continuously differentiable, then \( M \) is closed at \( (x, d) \). Explain how you can use this test in conjunction with the quadratic-fit method described in Section 8.3.

[8.9] Consider the problem to minimize \( 3\lambda - 2\lambda^2 + \lambda^3 + 2\lambda^4 \) subject to \( \lambda \geq 0 \).

a. Write a necessary condition for a minimum. Can you make use of this condition to find the global minimum?

b. Is the function strictly quasiconvex over the region \( \{\lambda : \lambda \geq 0\} \)?

Apply the Fibonacci search method to find the minimum.

c. Apply both the bisection search method and Newton's method to the above problem, starting from \( \lambda_1 = 6 \).

[8.10] Consider the following definitions:

A function \( \Theta : \mathbb{R} \to \mathbb{R} \) to be minimized is said to be strongly unimodal over the interval \([a, b]\) if there exists a \( \bar{\lambda} \) that minimizes \( \Theta \) over the interval; and for any \( \lambda_1, \lambda_2 \in [a, b] \) such that \( \lambda_1 < \lambda_2 \), we have

\[
\lambda_2 \leq \bar{\lambda} \quad \text{implies that } \Theta(\lambda_1) > \Theta(\lambda_2) \\
\lambda_1 \geq \bar{\lambda} \quad \text{implies that } \Theta(\lambda_1) < \Theta(\lambda_2). 
\]

A function \( \Theta : \mathbb{R} \to \mathbb{R} \) to be minimized is said to be strictly unimodal over the interval \([a, b]\) if there exists a \( \bar{\lambda} \) that minimizes \( \Theta \) over the interval; and for \( \lambda_1, \lambda_2 \in [a, b] \) such that \( \Theta(\lambda_1) \neq \Theta(\bar{\lambda}), \Theta(\lambda_2) \neq \Theta(\bar{\lambda}) \), and \( \lambda_1 < \lambda_2 \), we have

\[
\lambda_2 \leq \bar{\lambda} \quad \text{implies that } \Theta(\lambda_1) > \Theta(\lambda_2) \\
\lambda_1 \geq \bar{\lambda} \quad \text{implies that } \Theta(\lambda_1) < \Theta(\lambda_2). 
\]

a. Show that if \( \Theta \) is strongly unimodal over \([a, b]\), then \( \Theta \) is strongly quasiconvex over \([a, b]\). Conversely, show that if \( \Theta \) is strongly quasiconvex over \([a, b]\) and has a minimum in this interval, then it is strongly unimodal over the interval.

b. Show that if \( \Theta \) is strictly unimodal and continuous over \([a, b]\), then \( \Theta \) is strictly quasiconvex over \([a, b]\). Conversely, show that if \( \Theta \) is strictly quasiconvex over \([a, b]\) and has a minimum in this interval, then it is strictly unimodal over this interval.
[8.11] Let \( \theta : \mathbb{R} \rightarrow \mathbb{R} \) and suppose that we have the three points \((\lambda_1, \theta_1), (\lambda_2, \theta_2), \) and \((\lambda_3, \theta_3)\), where \( \theta_j = \theta(\lambda_j) \) for \( j = 1, 2, 3 \). Show that the quadratic curve \( q \) passing through these points is given by

\[
q(\lambda) = \frac{\theta_1(\lambda - \lambda_2)(\lambda - \lambda_3) + \theta_2(\lambda - \lambda_1)(\lambda - \lambda_3) + \theta_3(\lambda - \lambda_1)(\lambda - \lambda_2)}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3) + (\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3) + (\lambda_3 - \lambda_1)(\lambda_3 - \lambda_2)}.
\]

Furthermore, show that the derivative of \( q \) vanishes at the point \( \bar{\lambda} \) given by

\[
\bar{\lambda} = \frac{1}{2} \frac{b_{23}\theta_1 + b_{31}\theta_2 + b_{12}\theta_3}{a_{23}\theta_1 + a_{31}\theta_2 + a_{12}\theta_3},
\]

where \( a_{ij} = \lambda_i - \lambda_j \) and \( b_{ij} = \lambda_i^2 - \lambda_j^2 \). Find the quadratic curve passing through the points \((1, 4), (3, 1), \) and \((4, 7)\), and compute \( \bar{\lambda} \). Show that if \((\lambda_1, \lambda_2, \lambda_3)\) satisfy the three-point pattern (TPP), then \( \lambda_1 < \bar{\lambda} < \lambda_3 \). Also:

a. Propose a method for finding \( \lambda_1, \lambda_2, \lambda_3 \) such that \( \lambda_1 < \lambda_2 < \lambda_3, \theta_1 \geq \theta_2, \) and \( \theta_2 \leq \theta_3 \).

b. Show that if \( \theta \) is strictly quasiconvex, then the new interval of uncertainty defined by the revised \( \lambda_1 \) and \( \lambda_3 \) of the quadratic-fit line search indeed contains the minimum.

c. Use the procedure described in this exercise to minimize \(-3\lambda - 2\lambda^2 + 2\lambda^3 + 3\lambda^4\) over \( \lambda \geq 0 \).

[8.12] Let \( \theta : \mathbb{R} \rightarrow \mathbb{R} \) be a continuous strictly quasiconvex function. Let \( 0 \leq \lambda_1 < \lambda_2 < \lambda_3 \) and denote \( \theta_j = \theta(\lambda_j) \) for \( j = 1, 2, 3 \).

a. If \( \theta_1 = \theta_2 = \theta_3 \), show that this common value coincides with the value of \( min \{\theta(\lambda) : \lambda \geq 0\} \).

b. Let \((\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^3\) represent a three-point pattern iterate generated by the quadratic-fit algorithm described in Section 8.3. Show that the function \( \bar{\theta}(\lambda_1, \lambda_2, \lambda_3) \equiv \theta(\lambda_1) + \theta(\lambda_2) + \theta(\lambda_3) \) is a continuous function that satisfies the descent property \( \bar{\theta}[(\lambda_1, \lambda_2, \lambda_3)_{new}] < \bar{\theta}(\lambda_1, \lambda_2, \lambda_3) \) whenever \( \theta_1, \theta_2, \) and \( \theta_3 \) are not all equal to each other.

[8.13] Let \( \theta \) be pseudoconvex and continuously twice differentiable. Consider the algorithm of Section 8.3 with the modification that in Case 3, when \( \bar{\lambda} = \lambda_2 \), we let \( \lambda_{new} = (\lambda_1, \lambda_2, \bar{\lambda}) \) if \( \theta'(\lambda_2) > 0 \), we let \( \lambda_{new} = (\lambda_2, \bar{\lambda}, \lambda_3) \) if \( \theta'(\lambda_2) < 0 \), and we stop if \( \theta'(\lambda_2) = 0 \). Accordingly, if \( \lambda_1, \lambda_2, \) and \( \lambda_3 \) are not all distinct, let them be said to satisfy the three-point pattern (TPP) whenever \( \theta'(\lambda_2) < 0 \) if \( \lambda_1 = \lambda_2 < \lambda_3, \theta'(\lambda_2) > 0 \) if \( \lambda_1 < \lambda_2 = \lambda_3, \) and \( \theta'(\lambda_2) = 0 \) and
\( \theta''(\lambda_2) \geq 0 \) if \( \lambda_1 = \lambda_2 = \lambda_3 \). With this modification, suppose that we use the quadratic interpolation algorithm of Section 8.3 applied to \( \theta \) given a starting TPP \((\lambda_1, \lambda_2, \lambda_3)\), where the quadratic fit matches the two function values and the derivative \( \theta'(\lambda_2) \) whenever two of the three points \( \lambda_1, \lambda_2, \lambda_3 \) are coincident, and where at any iteration, if \( \theta'(\lambda_2) = 0 \), we put \( \lambda^* = (\lambda_2, \lambda_2, \lambda_2) \) and terminate. Define the solution set \( \Omega = \{(\lambda, \lambda, \lambda) : \theta'(\lambda) = 0\} \).

a. Let \( A \) define the algorithmic map that produces \( \lambda_{\text{new}} \in A(\lambda_1, \lambda_2, \lambda_3) \). Show that \( A \) is closed.

b. Show that the function \( \tilde{\theta}(\lambda_1, \lambda_2, \lambda_3) = \theta(\lambda_1) + \theta(\lambda_2) + \theta(\lambda_3) \) is a continuous descent function that satisfies \( \tilde{\theta}(\lambda_{\text{new}}) < \tilde{\theta}(\lambda_1, \lambda_2, \lambda_3) \) if \( \theta'(\lambda_2) \neq 0 \).

c. Hence, show that the algorithm defined either terminates finitely or generates an infinite sequence whose accumulation points lie in \( \Omega \).

d. Comment on the convergence of the algorithm and the nature of the solution obtained if \( \theta \) is strictly quasiconvex and twice continuously differentiable.

[8.14] In Section 8.2 we described Newton’s method for finding a point where the derivative of a function vanishes.

a. Show how the method can be used to find a point where the value of a continuously differentiable function is equal to zero. Illustrate the method for \( \theta(\lambda) = 2\lambda^3 - \lambda \), starting from \( \lambda_1 = 5 \).

b. Will the method converge for any starting point? Prove or give a counterexample.

[8.15] Show how the line search procedures of Section 8.1 can be used to find a point where a given function assumes the value zero. Illustrate by the function \( \theta \) defined by \( \theta(\lambda) = 2\lambda^2 - 5\lambda + 3 \). (Hint: Consider the absolute value function \( \hat{\theta} = |\theta| \).)

[8.16] In Section 8.2 we discussed the bisection search method for finding a point where the derivative of a pseudoconvex function vanishes. Show how the method can be used to find a point where a function is equal to zero. Explicitly state the assumptions that the function needs to satisfy. Illustrate by the function \( \theta \) defined by \( \theta(\lambda) = 2\lambda^3 - \lambda \) defined on the interval \([0.5, 10.0]\).

[8.17] It can be verified that in Example 9.2.4, for a given value of \( \mu \), if \( x_\mu = (x_1, x_2)' \), then \( x_1 \) satisfies

\[
2(x_1 - 2)^3 + \frac{\mu x_1 (8x_1^2 - 6x_1 + 1)}{4 + \mu} = 0.
\]
For \( \mu = 1, 10, 100, \) and \( 1000, \) find the value of \( x_1 \) satisfying the above equation, using a suitable procedure.

**[8.18]** Consider applying the steepest descent method to minimize \( f(x) \) versus the application of this method to minimize \( F(x) = \| \nabla f(x) \|^2. \) Assuming that \( f \) is quadratic with a positive definite Hessian, compare the rates of convergence of the two schemes and, hence, justify why the equivalent minimization of \( F \) is not an attractive strategy.

**[8.19]** Show that as a function of \( K_k, \) the expression in Equation (8.14) is maximized when \( K_k^2 = \alpha^2. \)

**[8.20]** Solve the problem to maximize \( 3x_1 + x_2 + 6x_1x_2 - 2x_1^2 + 2x_2^2 \) by the method of Hooke and Jeeves.

**[8.21]** Let \( H \) be an \( n \times n, \) symmetric, positive definite matrix with condition number \( \alpha. \) Then the Kantorovich inequality asserts that for any \( x \in \mathbb{R}^n, \) we have

\[
\frac{(x'Hx)(x'H^{-1}x)}{(x'x)^2} \geq \frac{4\alpha}{(1 + \alpha)^2}.
\]

Justify this inequality, and use it to establish Equation (8.18).

**[8.22]** Consider the problem to minimize \( (3 - x_1)^2 + 7(x_2 - x_1^2)^2. \) Starting from the point \((0, 0), \) solve the problem by the following procedures:

a. The cyclic coordinate method.
b. The method of Hooke and Jeeves.
c. The method of Rosenbrock.
d. The method of Davidon–Fletcher–Powell.

**[8.23]** Consider the following problem:

\[
\text{Minimize } \sum_{i=2}^{n} [100(x_i - x_{i-1}^2)^2 + (1 - x_{i-1})^2].
\]

For values of \( n = 5, 10 \) and 50, and starting from the solution \( x^0 = (-1.2, 1.0, -1.2, 1.0, \ldots), \) solve this problem using each of the following methods. (Write subroutines for evaluating the objective function, its gradient, and for performing a line search via the quadratic-fit method, and then use these subroutines to compose codes for the following methods. Also, you could use the previous iteration's step length as the initial step for establishing a three-point pattern (TPP) for the current iteration. Present a summary of comparative results.)
Unconstrained Optimization

a. The method of Hooke and Jeeves (use the line search variant and the same termination criteria as for the other methods, to facilitate comparisons).
b. Rosenbrock's method (again, use the line search variant as in Part a).
c. The steepest descent method.
d. Newton's method.
e. Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method.
f. The conjugate gradient method of Hestenes and Stiefel.
g. The conjugate gradient method of Fletcher and Reeves.
h. The conjugate gradient method of Polyak and Ribiere.

[8.24] Consider the problem to minimize \((x_1 - x_2^3)^2 + 3(x_1 - x_2)^4\). Solve the problem using each of the following methods. Do the methods converge to the same point? If not, explain.
a. The cyclic coordinate method.
b. The method of Hooke and Jeeves.
c. The method of Rosenbrock.
d. The method of steepest descent.
e. The method of Fletcher and Reeves.
f. The method of Davidon-Fletcher-Powell.
g. The method of Broyden-Fletcher-Goldfarb-Shanno (BFGS).

[8.25] Consider the model \(y = \alpha + \beta x + \gamma x^2 + \varepsilon\), where \(x\) is the independent variable, \(y\) is the observed dependent variable, \(\alpha\), \(\beta\), and \(\gamma\) are unknown parameters, and \(\varepsilon\) is a random component representing the experimental error. The following table gives the values of \(x\) and the corresponding values of \(y\). Formulate the problem of finding the best estimates of \(\alpha\), \(\beta\), and \(\gamma\) as an unconstrained optimization problem by minimizing:
a. The sum of squared errors.
b. The sum of the absolute values of the errors.
c. The maximum absolute value of the error.

For each case, find \(\alpha\), \(\beta\), and \(\gamma\) by a suitable method.

<table>
<thead>
<tr>
<th>(x)</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>3</td>
<td>3</td>
<td>-10</td>
<td>-25</td>
<td>-50</td>
<td>-100</td>
</tr>
</tbody>
</table>

[8.26] Consider the following problem:

Minimize \(2x_1 + x_2\)

subject to \(x_1^2 + x_2^2 = 9\)

\(-2x_1 - 3x_2 \leq 6.\)

a. Formulate the Lagrangian dual problem by incorporating both constraints into the objective function via the Lagrangian multipliers \(u_1\) and \(u_2\).
b. Using a suitable unconstrained optimization method, compute the
   gradient of the dual function $\theta$ at the point $(1, 2)$.

c. Starting from the point $\bar{u} = (1, 2)^T$, perform one iteration of the steepest
   ascent method for the dual problem. In particular, solve the following
   problem, where $d = \nabla \theta(\bar{u})$:

   \[
   \begin{align*}
   \text{Maximize } & \quad \theta(\bar{u} + \lambda d) \\
   \text{subject to } & \quad \bar{u}_2 + \lambda d_2 \geq 0 \\
   & \quad \lambda \geq 0.
   \end{align*}
   \]

[8.27] Let $f: R^n \to R$ be differentiable at $x$ and let the vectors $d_1, ..., d_n$ in $R^n$
   be linearly independent. Suppose that the minimum of $f(x + \lambda d_j)$ over $\lambda \in R$
   occurs at $\lambda = 0$ for $j = 1, ..., n$. Show that $\nabla f(x) = 0$. Does this imply that $f$ has a
   local minimum at $x$?

[8.28] Let $H$ be a symmetric $n \times n$ matrix, and let $d_1, ..., d_n$ be a set of
   characteristic vectors of $H$. Show that $d_1, ..., d_n$ are $H$-conjugate.

[8.29] Consider the problem in Equation (8.24) and suppose that $\varepsilon_k \geq 0$ is such
   that $H(x_k) + \varepsilon_k I$ is positive definite. Let $\Delta_k = -[H(x_k) + \varepsilon_k I]^{-1}\nabla f(x_k)$.
   Show that $\delta = x_{k+1} - x_k$, given by (8.22), and the Lagrange multiplier $\lambda = \varepsilon_k$
   satisfy the saddle point optimality conditions for (8.24). Hence, comment on the
   relationship between the Levenberg–Marquardt and trust region methods. Also
   comment on the case $\varepsilon_k = 0$.

[8.30] The following method for generating a set of conjugate directions for
   minimizing $f: R^n \to R$ is due to Zangwill [1967b]:

   **Initialization Step** Choose a termination scalar $\varepsilon > 0$, and choose an
   initial point $x_1$. Let $y_1 = x_1$, let $d_1 = -\nabla f(y_1)$, let $k = j = 1$, and go to the
   Main Step.

   **Main Step**

   1. Let $\lambda_j$ be an optimal solution to the problem to minimize $f(y_j + \lambda d_j)$
      subject to $\lambda \in R$, and let $y_{j+1} = y_j + \lambda_j d_j$. If $j = n$, go to
      Step 4; otherwise, go to Step 2.

   2. Let $d = -\nabla f(y_{j+1})$, and let $\hat{\mu}$ be an optimal solution to the
      problem to minimize $f(y_{j+1} + \mu d)$ subject to $\mu \geq 0$. Let $z_1 = y_{j+1}$
      + $\hat{\mu} d$. Let $i = 1$, and go to Step 3.

   3. If $\|\nabla f(z_i)\| < \varepsilon$, stop with $z_i$. Otherwise, let $\mu_i$ be an optimal
      solution to the problem to minimize $f(z_i + \mu d_i)$ subject to $\mu \in R$. 
Let $z_{i+1} = z_i + \mu_id_i$. If $i < j$, replace $i$ by $i + 1$, and repeat Step 3. Otherwise, let $d_{j+1} = z_{j+1} - y_{j+1}$, replace $j$ by $j + 1$, and go to Step 1.

4. Let $y_1 = x_k + 1 = y_{n+1}$. Let $d_1 = -\nabla f(y_1)$, replace $k$ by $k + 1$, let $j = 1$, and go to Step 1.

Note that the steepest descent search in Step 2 is used to ensure that $z_1 - y_1 \notin L(d_1, ..., d_j)$ for the quadratic case so that finite convergence is guaranteed.

Illustrate using the problem to minimize $(x_1 - 2)^4 + (x_1 - 2x_2)^2$, starting from the point $(0.0, 3.0)$.

[8.31] Suppose that $f$ is continuously twice differentiable and that the Hessian matrix is invertible everywhere. Given $x_k$, let $x_{k+1} = x_k + \lambda_k d_k$, where $d_k = -H(x_k)^{-1} \nabla f(x_k)$ and $\lambda_k$ is an optimal solution to the problem to minimize $f(x_k + \lambda d_k)$ subject to $\lambda \in R$. Show that this modification of Newton's method converges to a point in the solution set $\Omega = \{ \bar{x} : \nabla f(\bar{x})'H(\bar{x})^{-1}\nabla f(\bar{x}) = 0 \}$. Illustrate by minimizing $(x_1 - 2)^4 + (x_1 - 2x_2)^2$ starting from the point $(-2, 3)$.

[8.32] Let $a_1, ..., a_n$ be a set of linearly independent vectors in $R^n$, and let $H$ be an $n \times n$ symmetric positive definite matrix.

a. Show that the vectors $d_1, ..., d_n$ defined below are $H$-conjugate.

$$d_k = \begin{cases} a_k & \text{if } k = 1 \\ a_k - \sum_{i=1}^{k-1} \left( \frac{d_i'H a_k}{d_i'H d_i} \right) d_i & \text{if } k \geq 2. \end{cases}$$

b. Suppose that $a_1, ..., a_n$ are the unit vectors in $R^n$, and let $D$ be the matrix whose columns are the vectors $d_1, ..., d_n$ defined in Part a. Show that $D$ is upper triangular with all diagonal elements equal to 1.

c. Illustrate by letting $a_1 = (1, 0, 0)'$, $a_2 = (1, -1, 4)'$, $a_3 = (2, -1, 6)'$, and

$$H = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 3 & 2 \\ -1 & 2 & 2 \end{bmatrix}.$$ 

d. Illustrate by letting $a_1$, $a_2$, and $a_3$ be the unit vectors in $R^3$ and $H$ as given in Part c.

[8.33] Consider the following problem:
Minimize $2x_1^2 + 3x_1x_2 + 4x_2^2 + 2x_3^2 - 2x_2x_3 + 5x_1 + 3x_2 - 4x_3$.

Using Exercise 8.32 or any other method, generate three conjugate directions. Starting from the origin, solve the problem by minimizing along these directions.

[8.34] Show that analogous to (8.66), assuming exact line searches, a memoryless quasi-Newton update performed on a member of the Broyden family (taking $D_j = I$) results in a direction $d_{j+1} = -D_{j+1}Vf(y_{j+1})$, where

$$D_{j+1} = I - (1 - \phi) \frac{q_j q^T_j}{q^T_j q_j} - \phi \frac{p_j q^T_j}{q^T_j q_j}.$$ 

Observe that the equivalence with conjugate gradient methods results only when $\phi = 1$ (BFGS update).

[8.35] Show that there exists a value of $\phi$ [as given by Equation (8.50)] for the Broyden correction formula (8.46) that will yield $d_{j+1} = -D_{j+1}Vf(y_{j+1}) = 0.$

[Hint: Use $p_j = \lambda_j d_j - \lambda_j D_j Vf(y_j)$, $q_j = Vf(y_{j+1}) - Vf(y_j)$, and $d^T_j Vf(y_{j+1}) = p^T_j Vf(y_{j+1}) = Vf(y_{j+1}) D_j Vf(y_{j+1}) = 0.]$

[8.36] Use two sequential applications of the Sherman–Morrison–Woodbury formula given in Equation (8.55) to verify the inverse relationship (8.54) between (8.48) and (8.53).

[8.37] Derive the Hessian correction (8.53) for the BFGS update directly, following the scheme used for the update of the Hessian inverse via (8.41)–(8.45).

[8.38] Referring to Figure 8.24 and the associated discussion, verify that the minimization of the quadratic function $f$ from $y_j$ along the pattern direction $d_p = y'_{j+1} - y_j$ will produce the point $y_{j+2}$. [Hint: Let $y'_{j+2}$ denote the point thus obtained. Using the fact that $\nabla f(y'_{j+1}) \nabla f(y_{j+1}) = 0$ and that $\nabla f$ is linear, since $f$ is quadratic, show that $\nabla f(y'_{j+2})$ is orthogonal to both $\nabla f(y_{j+1})$ and to $d_p$, so that $y'_{j+2}$ is a minimizing point in the plane of Figure 8.24. Using Part 3 of Theorem 8.8.3, argue now that $y'_{j+2} = y_{j+2}$.]

[8.39] Consider the quadratic form $f(x) = c^T x + (1/2)x^T H x$, where $H$ is a symmetric $n \times n$ matrix. In many applications it is desirable to obtain separability in the variables by eliminating the cross-product terms. This could be done by rotating the axes as follows. Let $D$ be an $n \times n$ matrix whose columns $d_1, \ldots, d_n$ are $H$-conjugate. Letting $x = Dy$, verify that the quadratic form is
equivalent to $\sum_{j=1}^{n} \alpha_j y_j + (1/2)\sum_{j=1}^{n} \beta_j y_j^2$, where $(\alpha_1, ..., \alpha_n) = c^T D$, and $\beta_j = d_j^T H d_j$ for $j = 1, ..., n$. Furthermore, translating and rotating the axes could be accomplished by the transformation $x = Dy + z$, where $z$ is any vector satisfying $Hz + c = 0$, that is, $Vf(z) = 0$. In this case show that the quadratic form is equivalent to $[c^T z + (1/2)z^T Hz] + (1/2)\sum_{j=1}^{n} \beta_j y_j^2$. Use the result of this exercise to draw accurate contours of the quadratic form $3x_1 - 6x_2 + 2x_1^2 + x_1x_2 + 2x_2^2$.

**[8.40]** Consider the problem to maximize $-2x_1^2 - 3x_2^2 + 3x_1x_2 - 2x_1 + 4x_2$. Starting from the origin, solve the problem by the Davidon-Fletcher-Powell method, with $D_1$ as the identity. Also solve the problem by the Fletcher and Reeves conjugate gradient method. Note that the two procedures generate identical sets of directions. Show that, in general, if $D_1 = I$, then the two methods are identical for quadratic functions.

**[8.41]** Derive a quasi-Newton correction matrix $C$ for the Hessian approximation $B_k$ that achieves the minimum Frobenius norm (squared) $\sum_i \sum_j C_{ij}^2$, where $C_{ij}$ are the elements of $C$ (to be determined), subject to the quasi-Newton condition $(C + B_k)p_k = q_k$ and the symmetry condition $C = C^T$. [Hint: Set up the corresponding optimization problem after enforcing symmetry, and use the KKT conditions. This gives the Powell-Symmetric Broyden (PSB) update.]

**[8.42]** Solve the problem to minimize $2x_1 + 3x_2^2 + e^{2x_1^2 - x_2^2}$, starting with the point $(1, 0)$ and using both the Fletcher and Reeves conjugate gradient method and the BFGS quasi-Newton method.

**[8.43]** A problem of the following structure frequently arises in the context of solving a more general nonlinear programming problem:

Minimize $f(x)$
subject to $a_i \leq x_i \leq b_i$ for $i = 1, ..., m$.

a. Investigate appropriate modifications of the unconstrained optimization methods discussed in this chapter so that lower and upper bounds on the variables could be handled.

b. Use the results of Part a to solve the following problem:

Minimize $(x_1 - 2)^4 + (x_1 - 2x_2)^2$
subject to $4 \leq x_1 \leq 6$
$3 \leq x_2 \leq 5$. 
Consider the system of simultaneous equations

\[ h_i(x) = 0 \quad \text{for } i = 1, \ldots, \ell. \]

a. Show how to solve the above system by unconstrained optimization techniques. \([Hint: \text{Consider the problem to minimize } \sum_{i=1}^\ell |h_i(x)|^p, \text{ where } p \text{ is a positive integer.}]\]

b. Solve the following system:

\[
\begin{align*}
2(2x_1 - 2)^4 + (2x_1 - x_2)^2 - 4 &= 0 \\
x_1^2 - 2x_2 + 1 &= 0.
\end{align*}
\]

Consider the problem to minimize \( f(x) \) subject to \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \). A point \( x \) is said to be a KKT point if there exists a vector \( v \in \mathbb{R}^\ell \) such that

\[
\nabla f(x) + \sum_{i=1}^\ell v_i \nabla h_i(x) = 0
\]

\[ h_i(x) = 0 \quad \text{for } i = 1, \ldots, \ell. \]

a. Show how to solve the above system using unconstrained optimization techniques. \([Hint: \text{See Exercise 8.44.}]\)

b. Find the KKT point for the following problem:

\[
\begin{align*}
\text{Minimize} & \quad (x_1 - 3)^4 + (x_1 - 3x_2)^2 \\
\text{subject to} & \quad 2x_1^2 - x_2 = 0.
\end{align*}
\]

Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \).

a. Show that the KKT conditions are satisfied at a point \( x \) if there exist \( u_i \) and \( s_i \) for \( i = 1, \ldots, m \) such that

\[
\nabla f(x) + \sum_{i=1}^m u_i^2 \nabla g_i(x) = 0
\]

\[ g_i(x) + s_i^2 = 0 \quad \text{for } i = 1, \ldots, m \]

\[ u_i s_i = 0 \quad \text{for } i = 1, \ldots, m. \]

b. Show that unconstrained optimization techniques could be used to find a solution to the above system. \([Hint: \text{See Exercise 8.44.}]\)

c. Use a suitable unconstrained optimization technique to find a KKT point to the following problem:

\[
\begin{align*}
\text{Minimize} & \quad 3x_1^2 + 2x_2^2 - 2x_1 x_2 + 4x_1 + 6x_2 \\
\text{subject to} & \quad -2x_1 - 3x_2 + 6 \leq 0.
\end{align*}
\]
Consider the problem to minimize $x_1^2 + x_2^2$ subject to $x_1 + x_2 - 4 = 0$.

a. Find the optimal solution to this problem, and verify optimality by the KKT conditions.

b. One approach to solving the problem is to transform it into a problem of the form to minimize $x_1^2 + x_2^2 + \mu(x_1 + x_2 - 4)^2$, where $\mu > 0$ is a large scalar. Solve the unconstrained problem for $\mu = 10$ by a conjugate gradient method, starting from the origin.

Using induction, show that the inclusion of the extra term $\gamma_j d_1$ in Equation (8.68b), where $\gamma_j$ is as given therein, ensures the mutual H-conjugacy of the directions $d_1, \ldots, d_n$ thus generated.

Let $H$ be an $n \times n$ symmetric matrix, and let $f(x) = c^T x + (1/2)x^T H x$. Consider the following rank-one correction algorithm for minimizing $f$. First, let $D_1$ be an $n \times n$ positive definite symmetric matrix, and let $x_1$ be a given vector. For $j = 1, \ldots, n$, let $\lambda_j$ be an optimal solution to the problem to minimize $f(x_j + \lambda d_j)$ subject to $\lambda \in \mathbb{R}$, and let $x_{j+1} = x_j + \lambda_j d_j$, where $d_j = -D_j \nabla f(x_j)$ and $D_{j+1}$ is given by

\[
D_{j+1} = D_j + \frac{(p_j - D_j q_j)(p_j - D_j q_j)^T}{q_j'(p_j - D_j q_j)}
\]
\[
p_j = x_{j+1} - x_j
\]
\[
q_j = H p_j.
\]

a. Verify that the matrix added to $D_j$ to obtain $D_{j+1}$ is of rank 1.

b. For $j = 1, \ldots, n$, show that $p_i = D_{j+1} q_i$ for $i \leq j$.

c. Supposing that $H$ is invertible, does $D_{n+1} = H^{-1}$ hold?

d. Even if $D_j$ is positive definite, show that $D_{j+1}$ is not necessarily positive definite. This explains why a line search over the entire real line is used.

e. Are the directions $d_1, \ldots, d_n$ necessarily conjugate?

f. Use the above algorithm for minimizing $x_1 - 4x_2 + 2x_1^2 + 2x_1x_2 + 3x_2^2$.

g. Suppose that $q_j$ is replaced by $\nabla f(x_{j+1}) - \nabla f(x_j)$. Develop a procedure similar to that of Davidon–Fletcher–Powell for minimizing a nonquadratic function, using the above scheme for updating $D_j$. Use the procedure to minimize $(x_1 - 2)^4 + (x_1 - 2x_2)^2$. 

Chapter 8

[8.50] Consider the design of a conjugate gradient method in which \( d_{j+1} = -\nabla f(y_{j+1}) + \alpha_j d_j \) in the usual notation, and where, for a choice of a scale parameter \( s_{j+1} \), we would like \( s_{j+1}d_{j+1} \) to coincide with the Newton direction \(-H^{-1}\nabla f(y_{j+1})\), if at all possible. Equating \( s_{j+1}[-\nabla f(y_{j+1}) + \alpha_j d_j] = -H^{-1}\nabla f(y_{j+1})\), transpose both sides and multiply these by \( H d_j \), and use the quasi-Newton condition \( A_j H d = q \) to derive

\[
\alpha_j = \frac{\nabla f(y_{j+1})'q_j - (1/s_{j+1})\nabla f(y_{j+1})'p_j}{d_j'q_j}.
\]

a. Show that with exact line searches, the choice of \( s_{j+1} \) is immaterial. Moreover, show that as \( s_{j+1} \to \infty \), \( \alpha_j \to \alpha_j^{HS} \) in (8.57). Motivate the choice of \( s_{j+1} \) by considering the situation in which the Newton direction \(-H^{-1}\nabla f(y_{j+1})\) is, indeed, contained in the cone spanned by \(-\nabla f(y_{j+1})\) and \( d_j \) but is not coincident with \( d_j \). Hence, suggest a scheme for choosing a value for \( s_{j+1} \).

b. Illustrate, using Example 8.8.2, by assuming that at the previous iteration, \( y_j = (-1/2, 1)' \), \( d_j = (1, 0)' \), \( \lambda_j = 1/2 \) (inexact step), so that \( y_{j+1} = (0, 1)' \), and consider your suggested choice along with the choices (i) \( s_{j+1} = \infty \), (ii) \( s_{j+1} = 1 \), and (iii) \( s_{j+1} = 1/4 \) at the next iteration. Obtain the corresponding directions \( d_{j+1} = -\nabla f(y_{j+1}) + \alpha_j d_j \). Which of these can potentially lead to optimality? (Choice (ii) is Perry’s [1978] choice. Sherali and Ulular [1990] suggest the scaled version, prescribing a choice for \( s_{j+1} \).

[8.51] In this exercise we describe a modification of the simplex method of Spendley et al. [1962] for solving a problem of the form to minimize \( f(x) \) subject to \( x \in \mathbb{R}^n \). The version of the method described here is credited to Nelder and Mead [1965].

**Initialization Step** Choose the points \( x_1, x_2, \ldots, x_{n+1} \) to form a simplex in \( \mathbb{R}^n \). Choose a reflection coefficient \( \alpha > 0 \), an expansion coefficient \( \gamma > 1 \), and a positive contraction coefficient \( 0 < \beta < 1 \). Go to the Main Step.

**Main Step**

1. Let \( x_r, x_s \in \{x_1, \ldots, x_{n+1}\} \) be such that
Unconstrained Optimization

\[ f(x_r) = \min_{1 \leq j \leq n+1} f(x_j) \text{ and } f(x_s) = \max_{1 \leq j \leq n+1} f(x_j). \]

Let \( \bar{x} = \frac{1}{n+1} \sum_{j=1 \text{ or } j \neq s}^{n+1} x_j \), and go to Step 2.

2. Let \( \hat{x} = \bar{x} + \alpha(\bar{x} - x_s) \). If \( f(\hat{x}) > f(\hat{x}) \), let \( x_e = \bar{x} + \gamma(\hat{x} - \bar{x}) \), and go to Step 3. Otherwise, go to Step 4.

3. The point \( x_s \) is replaced by \( x_e \) if \( f(\hat{x}) > f(x_e) \) and by \( \hat{x} \) if \( f(\hat{x}) \leq f(x_e) \) to yield a new set of \( n+1 \) points. Go to Step 1.

4. If \( \max_{1 \leq j \leq n+1} \{f(x_j) : j \neq s\} \geq f(\hat{x}) \), then \( x_s \) is replaced by \( \hat{x} \) to form a new set of \( n+1 \) points, and we go to Step 1. Otherwise, go to Step 5.

5. Let \( x' \) be defined by \( f(x') = \min\{f(\hat{x}), f(x_s)\} \), and let \( x'' = \bar{x} + \beta(x' - \bar{x}) \). If \( f(x'') > f(x') \), replace \( x_j \) by \( x_j + (1/2)(x_r - x_j) \) for \( j = 1, \ldots, n+1 \), and go to Step 1. If \( f(x'') \leq f(x') \), then \( x'' \) replaces \( x_s \) to form a new set of \( n+1 \) points. Go to Step 1.

a. Let \( d_j \) be an \( n \)-vector with the \( j \)-th component equal to \( a \) and all other components equal to \( b \), where

\[ a = \frac{c}{n^{\sqrt{2}}}(\sqrt{n+1} + n - 1), \quad b = \frac{c}{n^{\sqrt{2}}}(\sqrt{n+1} - 1), \]

and where \( c \) is a positive scalar. Show that the initial simplex defined by \( x_1, \ldots, x_{n+1} \) could be chosen by letting \( x_{j+1} = x_1 + d_j \), where \( x_1 \) is selected arbitrarily. (In particular, show that \( x_{j+1} - x_1 \) for \( j = 1, \ldots, n \) are linearly independent. What is the interpretation of \( c \) in terms of the geometry of this initial simplex?)

b. Solve the problem to minimize \( 2x_1^2 + 2x_1x_2 + x_2^2 + 3x_2^2 - 3x_1 - 10x_3 \) using the simplex method described in this exercise.

[8.52] Consider the quadratic function \( f(y) = c^t y + (1/2)y^t H y \), where \( H \) is an \( n \times n \) symmetric, positive definite matrix. Suppose that we use some algorithm for which the iterate \( y_{j+1} = y_j - \lambda_j D_j \nabla f(y_j) \) is generated by an exact line search along the direction \( -D_j \nabla f(y_j) \) from the previous iterate \( y_j \), where \( D_j \) is some positive definite matrix. Then, if \( y^* \) is the minimizing solution for \( f \), and if \( e(y) = (1/2)(y - y^*)^t H(y - y^*) \) is an error function, show that at every step \( j \), we have

\[ e(y_{j+1}) \leq \frac{(\alpha_j - 1)^2}{(\alpha_j + 1)^2} e(y_j), \]
where $\alpha_j$ is the ratio of the largest to the smallest eigenvalue of $D_j H$.

[8.53] Consider the following method of parallel tangents credited to Shah et al. [1964] for minimizing a differentiable function $f$ of several variables:

**Initialization Step** Choose a termination scalar $\epsilon > 0$, and choose a starting point $x_1$. Let $y_0 = x_1$, $k = j = 1$, and go to the Main Step.

**Main Step**

1. Let $d = -\nabla f(x_k)$ and let $\hat{\lambda}$ be an optimal solution to the problem to minimize $f(x_k + \lambda d)$ subject to $\lambda \geq 0$. Let $y_1 = x_k + \hat{\lambda}d$. Go to Step 2.
2. Let $d = -\nabla f(y_j)$, and let $\lambda_j$ be an optimal solution to the problem to minimize $f(y_j + \lambda d)$ subject to $\lambda \geq 0$. Let $z_j = y_j + \lambda_j d$, and go to Step 3.
3. Let $d = z_j - y_{j-1}$, and let $\mu_j$ be an optimal solution to the problem to minimize $f(z_j + \mu d)$ subject to $\mu \in \mathbb{R}$. Let $y_{j+1} = z_j + \mu_j d$. If $j < n$, replace $j$ by $j + 1$, and go to Step 2. If $j = n$, go to Step 4.
4. Let $x_{k+1} = y_{n+1}$. If $\|x_{k+1} - x_k\| < \epsilon$, stop. Otherwise, let $y_0 = x_{k+1}$, replace $k$ by $k + 1$, let $j = 1$, and go to Step 1.

Using Theorem 7.3.4, show that the method converges. Solve the following problems using the method of parallel tangents:

a. Minimize $2x_1^2 + 3x_2^2 + 2x_1x_2 - 2x_1 - 6x_2$.

b. Minimize $x_1^2 + x_2^2 - 2x_1x_2 - 2x_1 - x_2$. (Note that the optimal solution for this problem is unbounded.)

c. Minimize $(x_1 - 3)^2 + (x_1 - 3x_2)^2$.

[8.54] Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable. Consider the following procedure for minimizing $f$:

**Initialization Step** Choose a termination scalar $\epsilon > 0$ and an initial step size $\Delta > 0$. Let $m$ be a positive integer denoting the number of allowable failures before reducing the step size. Let $x_1$ be the starting point and let the current upper bound on the optimal objective value be $UB = f(x_1)$. Let $\nu = 0$, let $k = 1$, and go to the Main Step.

**Main Step**

1. Let $d_k = -\nabla f(x_k)$, and let $x_{k+1} = x_k + \Delta d_k$. If $f(x_{k+1}) < UB$, let $\nu = 0$, $\hat{x} = x_{k+1}$, $UB = f(\hat{x})$, and go to Step 2. If, on the other hand, $f(x_{k+1}) \geq UB$, replace $\nu$ by $\nu + 1$. If $\nu = m$, go to Step 3; and if $\nu < m$, go to Step 2.
2. Replace \( k \) by \( k + 1 \), and go to Step 1.
3. Replace \( k \) by \( k + 1 \). If \( \Delta < \varepsilon \), stop with \( \hat{x} \) as an estimate of the optimal solution. Otherwise, replace \( \Delta \) by \( \Delta/2 \), let \( v = 0 \), let \( x_k = \hat{x} \), and go to Step 1.
   a. Can you prove convergence of the above algorithm for \( \varepsilon = 0 \)?
   b. Apply the above algorithm for the three problems in Exercise 8.53.

\[8.55\] The method of Rosenbrock can be described by the map \( A: \mathbb{R}^n \times U \times \mathbb{R}^n \to \mathbb{R}^n \times U \times \mathbb{R}^n \). Here \( U = \{ D : D \) is an \( n \times n \) matrix satisfying \( D'D = I \} \).

The algorithmic map \( A \) operates on the triple \((x, D, \lambda)\), where \( x \) is the current vector, \( D \) is the \( n \times n \) matrix whose columns are the directions of the previous iteration, and \( \lambda \) is the vector whose components \( \lambda_1, \ldots, \lambda_n \) give the distances moved along the directions \( d_1, \ldots, d_n \). The map \( A = A_3A_2A_1 \) is a composite map whose components are discussed in detail below.

1. \( A_1 \) is the point-to-point map defined by \( A_1(x, D, \lambda) = (x, \bar{D}) \), where \( \bar{D} \) is the matrix whose columns are the new directions defined by \( (8.9) \).

2. The point-to-set map \( A_2 \) is defined by \( (x, y, \bar{D}) \in A_2(x, \bar{D}) \) if minimizing \( f \), starting from \( x \), in the directions \( \bar{d}_1, \ldots, \bar{d}_n \) leads to \( y \). By Theorem 7.3.5, the map \( A_2 \) is closed.

3. \( A_3 \) is the point-to-point map defined by \( A_3(x, y, \bar{D}) = (y, \bar{D}, \bar{\lambda}) \), where \( \bar{\lambda} = (\bar{D})^{-1}(y - x) \).
   a. Show that the map \( A_1 \) is closed at \((x, D, \lambda)\) if \( \lambda_j \neq 0 \) for \( j = 1, \ldots, n \).
   b. Is the map \( A_1 \) closed if \( \lambda_j = 0 \) for some \( j \)? (Hint: Consider the sequence \( D_k = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \) and \( \lambda_k = \begin{bmatrix} 1/k \\ 1 \end{bmatrix} \).
   c. Show that \( A_3 \) is closed.
   d. Verify that the function \( f \) could be used as a descent function.
   e. Discuss the applicability of Theorem 7.2.3 to prove convergence of Rosenbrock's procedure. (This exercise illustrates that some difficulties could arise in viewing the algorithmic map as a composition of several maps. In Section 8.5 a proof of convergence was provided without decomposing the map \( A \).)

\[8.56\] Consider the problem to minimize \( f(x) \) subject to \( x \in \mathbb{R}^n \), and consider the following algorithm credited to Powell [1964] (and modified by Zangwill [1967b] as in Part c).

**Initialization Step** Choose a termination scalar \( \varepsilon > 0 \). Choose an initial point \( x_1 \), let \( d_1, \ldots, d_n \) be the coordinate directions, and let \( k = j = i = 1 \). Let \( z_1 = y_1 = x_1 \), and go to the Main Step.
Main Step

1. Let $\lambda_i$ be an optimal solution to the problem to minimize $f(z_i + \lambda d_i)$ subject to $\lambda \in \mathbb{R}$, and let $z_{i+1} = z_i + \lambda_i d_i$. If $i < n$, replace $i$ by $i + 1$, and repeat Step 1. Otherwise, go to Step 2.

2. Let $d = z_{n+1} - z_1$, and let $\hat{\lambda}$ be an optimal solution to the problem to minimize $f(z_{n+1} + \lambda d)$ subject to $\lambda \in \mathbb{R}$. Let $y_{j+1} = z_{n+1} + \hat{\lambda} d$. If $j < n$, replace $d_\ell$ by $d_\ell = d_{\ell+1}$ for $\ell = 1, \ldots, n - 1$, let $d_n = d$, let $z_1 = y_{j+1}$, let $i = 1$, replace $j$ by $j + 1$, and go to Step 1. Otherwise, $j = n$, and go to Step 3.

3. Let $x_{k+1} = y_{n+1}$. If $\|x_{k+1} - x_k\| < \varepsilon$, stop. Otherwise, let $i = j = 1$, let $z_1 = y_1 = x_{k+1}$, replace $k$ by $k + 1$, and go to Step 1.

a. Suppose that $f(x) = c^T x + (1/2)x^T H x$, where $H$ is an $n \times n$ symmetric matrix. After one pass through the main step, show that if $d_1, \ldots, d_n$ are linearly independent, then they are also $H$-conjugate, so that by Theorem 8.8.3, an optimal solution is produced in one iteration.

b. Consider the following problem credited to Zangwill [1967b]:

\[
\text{Minimize } (x_1 - x_2 + x_3)^2 + (-x_1 + x_2 + x_3)^2 + (x_1 + x_2 - x_3)^2.
\]

Apply Powell's method discussed in this exercise, starting from the point $(1/2, 1, 1/2)$. Note that the procedure generates a set of dependent directions and hence will not yield the optimal point $(0, 0, 0)$.

c. Zangwill [1967b] proposed a slight modification of Powell's method to guarantee linear independence of the direction vectors. In particular, in Step 2, the point $z_1$ is obtained from $y_{j+1}$ by a spacer step application, such as one iteration of the cyclic coordinate method. Show that this modification indeed guarantees linear independence, and hence, by Part a, finite convergence for a quadratic function is assured.

d. Apply Zangwill's modified method to solve the problem of Part b.

e. If the function is not quadratic, consider the introduction of a spacer step so that in Step 3, $z_1 = y_1$ is obtained by the application of one iteration of the cyclic coordinate method starting from $x_{k+1}$. Use Theorem 7.3.4 to prove convergence.

[8.57] Solve the Lagrangian dual problem of Example 6.4.1 using the subgradient algorithm. Resolve using the deflected subgradient strategy suggested in Section 8.9.

[8.58] Consider the problem of finding $\bar{x} = P_G(x)$, where $G = \{y : \xi_j^T y \leq \beta_j, \text{ for } j = 1, 2\}$.

a. Formulate this as a linearly constrained quadratic optimization problem and write the KKT conditions for this problem. Explain why these KKT
conditions are both necessary and sufficient for optimality for this problem.

b. Prescribe a closed-form solution to these conditions, enumerating cases as necessary. Illustrate geometrically each such case identified.

c. Identify the above analysis with the main computation in the Polyak–Kelly cutting plane algorithm as embodied by Equations (8.80) and (8.81).

[8.59] Solve the example of Exercise 6.30 using the subgradient optimization algorithm starting with the point \((0,4)\). Re-solve using the deflected subgradient strategy suggested in Section 8.9.

[8.60] Consider the problem of finding the projection \(\mathbf{x}^* = P_X(\bar{x})\) of the point \(\bar{x}\) onto \(X = \{x : \alpha^T x = \beta, \ell \leq x \leq u\}\), where \(x, \bar{x}, x^* \in \mathbb{R}^n\). The following variable dimension algorithm projects the current point successively onto the equality constraint and reduces the problem to an equivalent one in a lower-dimensional space, or else stops. Justify the various steps of this algorithm. Illustrate by projecting the point \((-2, 3, 1, 2)'\) onto \(\{x : x_1 + x_2 + x_3 + x_4 = 1, 0 \leq x_i \leq 1\ \text{for all} \ i\}\). (This method is a generalization of the procedures that appear in Bitran and Hax [1976] and in Sherali and Shetty [1980b].)

Initialization Set \((\bar{x}^0, \ell^0, u^0, \beta^0) = (\bar{x}, \ell, u, \beta)\), where \(I = \{i : \alpha_i \neq 0\}\). For \(i \in I\), put \(x_i^* = \bar{x}_i\) if \(\ell_i \leq \bar{x}_i \leq u_i\), \(x_i^* = \ell_i\) if \(\bar{x}_i < \ell_i\) and \(x_i^* = u_i\) if \(\bar{x}_i > u_i\). Let \(k = 0\).

Step 1 Compute the projection \(\hat{x}^k\) of \(\bar{x}^k\) onto the equality constraint in the subspace \(I^k\) according to

\[
\hat{x}_i^k = \bar{x}_i^k + \frac{\beta^k - \sum_{i \in I} \alpha_i \bar{x}_i^k}{\sum_{i \in I^k} \alpha_i^2} \alpha_i \quad \text{for each} \ i \in I^k.
\]

If \(\hat{x}_i^k \leq x_i^* \leq u_i^k\) for all \(i \in I^k\), put \(x_i^* = \hat{x}_i^k\) for all \(i \in I^k\), and stop. Otherwise, proceed to Step 2.

Step 2 Define \(J_1 = \{i \in I^k : \hat{x}_i^k \leq \ell_i^k\}\), \(J_2 = \{i \in I^k : \hat{x}_i^k \geq u_i^k\}\), and compute

\[
\gamma = \beta^k + \sum_{i \in J_1} \alpha_i (\ell_i^k - \hat{x}_i^k) + \sum_{i \in J_2} \alpha_i (u_i^k - \hat{x}_i^k).
\]

If \(\gamma = \beta^k\), then put \(x_i^* = \ell_i^k\) for \(i \in J_1\), \(x_i^* = u_i^k\) for \(i \in J_2\), and \(x_i^* = \hat{x}_i^k\) for \(i \in I^k - J_1 \cup J_2\), and stop. Otherwise, define
\[ J_3 = \{i \in J_1 : \alpha_i > 0\} \quad \text{and} \quad J_4 = \{i \in J_2 : \alpha_i < 0\} \quad \text{if} \ \gamma > \beta^k \]

\[ J_3 = \{i \in J_1 : \alpha_i < 0\} \quad \text{and} \quad J_4 = \{i \in J_2 : \alpha_i > 0\} \quad \text{if} \ \gamma < \beta^k. \]

Set \( x_i^* = \ell_i^k \) if \( i \in J_3 \), and \( x_i^* = u_i^k \) if \( i \in J_4 \). (Note: \( J_3 \cup J_4 \neq \emptyset \).) Update \( r^{k+1} = I^k - J_3 \cup J_4 \). If \( I^{k+1} = \emptyset \), then stop. Otherwise, update \( (x_i^k = \hat{x}_i^k) \) for \( i \in I^{k+1} \), \( \ell_i^{k+1} = \max \{ \ell_i^k, \hat{x}_i^k \} \) if \( \alpha_i(\beta^k - \gamma) > 0 \), and \( \ell_i^{k+1} = \ell_i^k \) otherwise, for \( i \in I^{k+1} \), \( u_i^{k+1} = \min \{ u_i^k, \hat{x}_i^k \} \) if \( \alpha_i(\beta^k - \gamma) < 0 \), and \( u_i^{k+1} = u_i^k \) otherwise, for \( i \in I^{k+1} \). And \( \beta^{k+1} = \beta^k - \sum_{i \in J_3} \alpha_i \ell_i^k - \sum_{i \in J_4} \alpha_i u_i^k \).

Increment \( k \) by 1 and go to Step 1.

**Notes and References**

We have discussed several iterative procedures for solving an unconstrained optimization problem. Most of the procedures involve a line search of the type described in Sections 8.1 through 8.3 and, by and large, the effectiveness of the search direction and the efficiency of the line search method greatly affect the overall performance of the solution technique. The Fibonacci search procedure discussed in Section 8.1 is credited to Kiefer [1953]. Several other search procedures, including the golden section method, are discussed in Wilde [1964] and Wilde and Beightler [1967]. These references also show that the Fibonacci search procedure is the best for unimodal functions in that it reduces the maximum interval of uncertainty with the least number of observations.

Another class of procedures uses curve fitting, as discussed in Section 8.3 and illustrated by Exercises 8.11 through 8.13. If a function \( f \) of one variable is to be minimized, the procedures involve finding an approximating quadratic or cubic function \( q \). In the quadratic case, the function is selected such that given three points \( \lambda_1, \lambda_2, \) and \( \lambda_3 \), the functional values of \( f \) and \( q \) are equal at these points. In the cubic case, given two points \( \lambda_1 \) and \( \lambda_2 \), \( q \) is selected such that the functional values and derivatives of both functions are the same at these points. In any case, the minimum of \( q \) is determined, and this point replaces one of the initial points. Refer to Davidon [1959], Fletcher and Powell [1963], Kowalik and Osborne [1968], Luenberger [1973a/1984], Pierre [1969], Powell [1964], and Swann [1964] for more detailed discussions, particularly on precautions to be taken to ensure convergence. Some limited computational studies on the efficiency of this approach may be found in Himmelblau [1972b] and Murtagh and Sargent [1970]. See Armijo [1966] and Luenberger [1973a/1984] for further discussions on inexact line searches.

Among the gradient-free methods, the method of Rosenbrock [1960], discussed in Section 8.4, and the method of Zangwill [1967b], discussed in Exercises 8.30 and 8.56, are generally considered quite efficient. As originally proposed, the Rosenbrock method and the procedure of Hooke and Jeeves [1961] do not use line searches but employ instead discrete steps along the search directions. Incorporating a line search within Rosenbrock's procedure
was suggested by Davies, Swann, and Campey and is discussed by Swann [1964]. An evaluation of this modification is presented by Fletcher [1965] and Box [1966].

There are yet other derivative-free methods for unconstrained minimization. A procedure that is distinctively different, called the sequential simplex search method, is described in Exercise 8.51. The method was proposed by Spendley et al. [1962] and modified by Nelder and Mead [1965]. The method essentially looks at the functional values at the extreme points of a simplex. The worst extreme point is rejected and replaced by a new point along the line joining this point and the centroid of the remaining points. The process is repeated until a suitable termination criterion is satisfied. Box [1966], Jacoby et al. [1972], Kowalik and Osborne [1968], and Parkinson and Hutchinson [1972a] compare this method with some of the other methods discussed earlier. Parkinson and Hutchinson [1972b] have presented a detailed analysis on the efficiency of the simplex method and its variants. The simplex method seems to be less effective as the dimensionality of the problem increases.

The method of steepest descent proposed by Cauchy in the middle of the nineteenth century continues to be the basis of several gradient-based solution procedures. For example, see Gonzaga [1990] for a polynomial-time, scaled steepest descent algorithm for linear programming. The method of steepest descent uses a first-order approximation of the function being minimized and usually performs poorly as the optimum is reached. On the other hand, Newton’s method uses a second-order approximation and usually performs well at points close to the optimum. In general, however, convergence is guaranteed only if the starting point is close to the solution point. For a discussion on Newton-Raphson methods, see Fletcher [1987]. Burns [1993] presents a powerful alternative for solving systems of signomial equations in variables that are restricted to be positive in value. Fletcher [1987] and Dennis and Schnabel [1983] give a good discussion of Levenberg [1944]-Marquardt [1963] methods used for implementing the modification of Newton’s method by replacing $H(x_k)$ by $H(x_k) + \varepsilon_k I$ to maintain positive definiteness, and also of the relationship of this to trust region methods. For a survey and implementation aspects, see Moré [1977]. For a discussion and survey on trust region methods and related convergence aspects, see Conn et al. [1988b, 1997, 2000] and Powell [2003]. Ye [1990] presents a method to solve the subproblems appearing in such methods. We also introduced the dog-leg trajectory of Powell [1970a], which compromises between the steepest descent step and Newton’s step (see also Dennis and Schnabel [1983] and Fletcher [1987]). For another scheme of combining the steepest descent and Newton’s methods, see Luenberger [1973a/1984]. Renegar [1988] presents a polynomial-time algorithm for linear programming based on Newton’s method. Burns [1989] gives some very interesting graphical plots to illustrate the convergence behavior of some of the aforementioned algorithms.

Among the unconstrained optimization techniques, methods using conjugate directions are considered efficient. For a quadratic function, these methods give the optimum in, at most, $n$ steps. Among the derivative-free methods of this type are the method of Zangwill, discussed in Exercises 8.30
and 8.56, the method of Powell, discussed in Exercise 8.56, and the PARTAN method credited to Shah et al. [1964], discussed in Exercise 8.53. Sorensen [1969] has shown that for quadratic functions, PARTAN is far less efficient than the conjugate gradient methods discussed in Section 8.8.

In yet another class, the direction of movement \( \mathbf{d} \) is taken to be \(-\mathbf{D} \nabla f(x)\), where \( \mathbf{D} \) is a positive definite matrix that approximates the inverse of the Hessian matrix. This class is usually referred to as quasi-Newton methods. (See Davidon et al. [1991] for a note on terminology in this area.) One of the early methods of minimizing a nonlinear function using this approach is that of Davidon [1959], which was simplified and reformulated by Fletcher and Powell [1963] and is referred to as the variable metric method. A useful generalization of the Davidon-Fletcher-Powell method was proposed by Broyden [1967]. Essentially, Broyden introduced a degree of freedom in updating the matrix \( \mathbf{D} \). A particular choice of this degree of freedom was then proposed by Broyden [1970], Fletcher [1970a], Goldfarb [1970], and Shanno [1970]. This has led to the well-known BFGS update technique. Gill et al. [1972], among several others, have shown that this modification performs more efficiently than the original method for most problems. For an extension on updating conjugate directions via the BFGS approach, see Powell [1987].

In 1972, Powell showed that the Davidon-Fletcher-Powell method converges to an optimal solution if the objective function is convex, if the second derivatives are continuous, and if an exact line search is used. Under stronger assumptions, Powell [1971b] showed that the method converges superlinearly. In 1973, Broyden et al., gave local convergence results for the case where the step size is fixed equal to 1 and proved superlinear convergence under certain conditions. Under suitable assumptions, Powell [1976] showed that a version of the variable metric method without exact line searches converges to an optimal solution if the objective function is convex. Furthermore, he showed that if the Hessian matrix is positive definite at the solution point, the rate of convergence is superlinear. For further reading on variable metric methods and their convergence characteristics, see Broyden et al. [1973], Dennis and Moré [1974], Dixon [1972a-e], Fletcher [1987], Gill and Murray [1974a,b], Greenstadt [1970], Huang [1970], and Powell [1971b, 1976, 1986, 1987].

The variable metric methods discussed above update the matrix \( \mathbf{D} \) by adding to it two matrices, each having rank 1, and hence, this class is also referred to as rank-two correction procedures. A slightly different strategy for estimating second derivatives is to update the matrix \( \mathbf{D} \) by adding to it a matrix of rank 1. This rank-one correction method was introduced briefly in Exercise 8.49. For further details of this procedure, see Broyden [1967], Davidon [1969], Dennis and Schnabel [1983], Fiacco and McCormick [1968], Fletcher [1987], and Powell [1970a]. Conn et al. [1991] provide a detailed convergence analysis.

Among the conjugate methods using gradient information, the method of Fletcher and Reeves generates conjugate directions by taking a suitable convex combination of the current gradient and the direction used at the previous iteration. The original idea presented by Hestenes and Stiefel [1952] led to the development of this method, as well as to the conjugate gradient algorithms of
Unconstrained Optimization


Several authors have attempted to use unconstrained optimization methods for solving nonlinear problems with constraints. Note that if an unconstrained optimization technique is extended to handle constraints by simply rejecting infeasible points during the search procedure, it would lead to premature termination. A successful and frequently used approach is to define an auxiliary unconstrained problem such that the solution of the unconstrained problem yields the solution of the constrained problem. This is discussed in detail in Chapter 9. A second approach is to use an unconstrained optimization method when we are in the interior of the feasible region and to use one of the suitable constrained optimization methods discussed in Chapter 10 when we are at a point on the boundary of the feasible region. Several authors have also modified the unconstrained optimization methods to handle constraints. Goldfarb [1969a] has extended the Davidon–Fletcher–Powell method to handle problems having linear constraints utilizing the concept of gradient projection. The method was generalized by Davies [1970] to handle nonlinear constraints. Coleman and Fenyes [1989] discuss and survey other quasi-Newton methods for solving equality constrained problems. Klingman and Himmelblau [1964] project the search direction in the method of Hooke and Jeeves into the intersection of binding constraints, which leads to a constrained version of the Hooke and Jeeves algorithm. Glass and Cooper [1965] have presented another constrained version of the Hooke and Jeeves algorithm. The method of
Rosenbrock has been extended by Davies and Swann [1969] to handle linear constraints. In Exercise 8.51 we discussed the simplex method for solving unconstrained problems. In 1965, Box extended this approach to constrained problems. For other alternative extensions of the simplex method, see Dixon [1973], Friedman and Pinder [1972], Ghani [1972], Guin [1968], Keefer [1973], Mitchell and Kaplan [1968], Paviani and Himmelblau [1969], and Umida and Ichikawa [1971].

For comprehensive surveys of different algorithms for solving unconstrained problems, refer to Bertsekas [1995], Dennis and Schnabel [1983], Fletcher [1969b, 1987], Gill et al. [1981], Nash and Sofer [1996], Nocedal and Wright [1999], Powell [1970b], Reklaitis and Phillips [1975], and Zoutendijk [1970a,b]. Furthermore, there are numerous studies reporting computational experience with different algorithms. Most of them study the effectiveness of the methods by solving relatively small test problems of different degrees of complexity. For discussions on the efficiency of the various unconstrained minimization algorithms, see Bard [1970], Cragg and Levy [1969], Fiacco and McCormick [1968], Fletcher [1987], Gill et al. [1981], Himmelblau [1972b], Huang and Levy [1970], Murtagh and Sargent [1970], and Sargent and Sebastian [1972]. Computer program listings of some of the algorithms may be found in Brent [1973] and Himmelblau [1972b]. The Computer Journal and the Journal of the ACM also publish computer listings of nonlinear programming algorithms. Also, the NEOS Server for Optimization (http://www.neos.mcs.anl.gov/) maintains state-of-the-art optimization software on their Web site, where users can solve various types of optimization problems.

Finally, in Section 8.9, we presented the essence of subgradient optimization techniques following Held et al. [1974] and Polyak [1967, 1969a] with choices of step sizes as in Bazaraa and Sherali [1981], Held et al. [1974], and Sherali and Ulular [1989]. See Allen et al. [1987] for a useful theoretical justification of various practical step-size rules. Barahona and Anbil [2000], Bazaraa and Goode [1979], Larsson et al. [1996, 2004], Lim and Sherali [2005b], Sherali and Lim [2004, 2005], and Sherali and Myers [1988], discuss various aspects of using subgradient optimization in the context of Lagrangian duality. Related recovery of primal solutions are discussed by Larsson et al. [1999] and Sherali and Choi [1996]. Borwein [1982] addresses the existence of subgradients, and Hiriart–Urruty [1978] discusses optimality conditions under nondifferentiability. Various schemes have been suggested to accelerate the convergence of subgradient methods, but they still require a fine tuning of parameters to yield an adequate performance, particularly for large problems. Among these, the simplest to implement, and suitable for large problems, are the conjugate subgradient methods discussed in Camerini et al. [1975], Sherali and Ulular [1989], and Wolfe [1976]. Other more efficient methods, but requiring more effort and storage and suitable for relatively smaller problems, are (1) the space dilation techniques of Sherali et al. [2001a] and Shor [1970, 1975, 1977b, 1985], particularly those using dilation in the direction of the difference between two successive subgradients that imitate variable metric methods (see also Minoux [1986] and Shor [1977a] for surveys); (2) the extension of Davidon methods to nondifferentiable problems as in Lemarechal [1975]; and (3) the
bundle algorithms described in Feltenmark and Kiwiel [2000], Kiwiel [1985, 1989, 1991, 1995], Lemarechal [1978, 1980], and Lemarechal and Mifflin [1978], which attempt to construct descent directions via approximations of subdifferentials. For an insightful connection between the symmetric rank-one quasi-Newton method and space dilation methods, see Todd [1986]. Goffin and Kiwiel [1999] and Sherali et al. [2000] discuss some effective variable target value methods, and Lim and Sherali [2005a,b] and Sherali and Lim [2004], provide convergence analyses and computational experience for several combinations of such methods with the different direction finding and line search techniques, including the average direction strategy of Sherali and Ulular [1989], the volume algorithm of Barahona and Anbil [2000], and the Polyak–Kelly cutting plane method of Polyak [1969a] and its modification by Sherali et al. [2001], and Lim and Sherali [2005a].
Chapter Penalty and Barrier Functions

In this chapter we discuss nonlinear programming problems having equality and inequality constraints. The approach used is to convert the problem into an equivalent unconstrained problem or into a problem having simple bound constraints, so that the algorithms developed in Chapter 8 can be used. However, in practice, a sequence of problems are solved because of computational considerations, as discussed later in the chapter. Basically, there are two alternative approaches. The first is called the penalty or the exterior penalty function method, in which a term is added to the objective function to penalize any violation of the constraints. This method generates a sequence of infeasible points (hence its name) whose limit is an optimal solution to the original problem. The second method is called the barrier or interior penalty function method, in which a barrier penalty term that prevents the points generated from leaving the feasible region is added to the objective function. This method generates a sequence of feasible points whose limit is an optimal solution to the original problem.

Following is an outline of the chapter.

Section 9.1: Concept of Penalty Functions The concept of penalty functions is introduced. A geometric interpretation of the method is also discussed.

Section 9.2: Exterior Penalty Function Methods The exterior penalty function methods are discussed in detail, and the main convergence theorem is developed. The method is illustrated by means of a numerical example. Computational difficulties associated with this class of algorithms are discussed along with related convergence rate aspects.

Section 9.3: Exact Absolute Value and Augmented Lagrangian Penalty Methods To alleviate the computational difficulties associated with having to take the penalty parameter to infinity in order to recover an optimal solution to the original problem, we introduce the concept of exact penalty functions. Both the absolute value $(\ell_1)$ and the augmented Lagrangian exact penalty function methods are discussed along with their related computational considerations.
9.1 Concept of Penalty Functions

Methods using penalty functions transform a constrained problem into a single unconstrained problem or into a sequence of unconstrained problems. The constraints are placed into the objective function via a penalty parameter in a way that penalizes any violation of the constraints. To motivate penalty functions, consider the following problem having the single constraint \( h(x) = 0 \):

\[
\text{Minimize } f(x) \\
\text{subject to } h(x) = 0.
\]

Suppose that this problem is replaced by the following unconstrained problem, where \( \mu > 0 \) is a large number:

\[
\text{Minimize } f(x) + \mu h^2(x) \\
\text{subject to } x \in \mathbb{R}^n.
\]

We can intuitively see that an optimal solution to the above problem must have \( h^2(x) \) close to zero, because otherwise, a large penalty \( \mu h^2(x) \) will be incurred.

Now consider the following problem having single inequality constraint \( g(x) \leq 0 \):

\[
\text{Minimize } f(x) \\
\text{subject to } g(x) \leq 0.
\]

It is clear that the form \( f(x) + \mu g^2(x) \) is not appropriate, since a penalty will be incurred whether \( g(x) < 0 \) or \( g(x) > 0 \). Needless to say, a penalty is desired only if the point \( x \) is not feasible, that is, if \( g(x) > 0 \). A suitable unconstrained problem is therefore given by:

\[
\text{Minimize } f(x) + \mu \text{ maximum } \{0, g(x)\} \\
\text{subject to } x \in \mathbb{R}^n.
\]
Penalty and Barrier Functions

If \( g(x) \leq 0 \), then \( \max\{0, g(x)\} = 0 \), and no penalty is incurred. On the other hand, if \( g(x) > 0 \), then \( \max\{0, g(x)\} > 0 \) and the penalty term \( \mu g(x) \) is realized.

However, observe that at points \( x \) where \( g(x) = 0 \), the foregoing objective function might not be differentiable, even though \( g \) is differentiable. If differentiability is desirable in such a case, we could, for example, consider instead a penalty function term of the type \( \mu \{\max\{0, g(x)\}\}^2 \).

In general, a suitable penalty function must incur a positive penalty for infeasible points and no penalty for feasible points. If the constraints are of the form \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) and \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \), a suitable penalty function \( \alpha \) is defined by

\[
\alpha(x) = \sum_{i=1}^{m} \phi(g_i(x)) + \sum_{i=1}^{\ell} \psi(h_i(x)),
\]

where \( \phi \) and \( \psi \) are continuous functions satisfying the following:

\[
\begin{align*}
\phi(y) &= 0 \quad \text{if } y \leq 0 \quad \text{and} \quad \phi(y) > 0 \quad \text{if } y > 0 \\
\psi(y) &= 0 \quad \text{if } y = 0 \quad \text{and} \quad \psi(y) > 0 \quad \text{if } y \neq 0.
\end{align*}
\]

Typically, \( \phi \) and \( \psi \) are of the forms

\[
\begin{align*}
\phi(y) &= [\max\{0, y\}]^p \\
\psi(y) &= |y|^p
\end{align*}
\]

where \( p \) is a positive integer. Thus, the penalty function \( \alpha \) is usually of the form

\[
\alpha(x) = \sum_{i=1}^{m} [\max\{0, g_i(x)\}]^p + \sum_{i=1}^{\ell} |h_i(x)|^p.
\]

We refer to the function \( f(x) + \mu \alpha(x) \) as the auxiliary function. Later, we introduce an augmented Lagrangian function in which the Lagrangian function [and not simply \( f(x) \)] is augmented with a penalty term.

9.1.1 Example

Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad x \\
\text{subject to} & \quad -x + 2 \leq 0.
\end{align*}
\]

Let \( \alpha(x) = [\max\{0, g(x)\}]^2 \). Then

\[
\alpha(x) = \begin{cases} 
0 & \text{if } x \geq 2 \\
(-x + 2)^2 & \text{if } x < 2.
\end{cases}
\]
Figure 9.1 shows the penalty and auxiliary functions $\alpha$ and $f + \mu \alpha$. Note that the minimum of $f + \mu \alpha$ occurs at the point $2 - (1/2 \mu)$ and approaches the minimizing point $\bar{x} = 2$ of the original problem as $\mu$ approaches $\infty$.

### 9.1.2 Example

Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad x_1^2 + x_2^2 \\
\text{subject to} & \quad x_1 + x_2 - 1 = 0.
\end{align*}
\]

The optimal solution lies at the point $(1/2, 1/2)$ and has objective value $1/2$.

Now, consider the following penalty problem, where $\mu > 0$ is a large number:

\[
\begin{align*}
\text{Minimize} & \quad x_1^2 + x_2^2 + \mu(x_1 + x_2 - 1)^2 \\
\text{subject to} & \quad (x_1, x_2) \in \mathbb{R}^2.
\end{align*}
\]

Note that for any $\mu \geq 0$, the objective function is convex. Thus, a necessary and sufficient condition for optimality is that the gradient of $x_1^2 + x_2^2 + \mu(x_1 + x_2 - 1)^2$ is equal to zero, yielding

\[
\begin{align*}
x_1 + \mu(x_1 + x_2 - 1) &= 0 \\
x_2 + \mu(x_1 + x_2 - 1) &= 0.
\end{align*}
\]

Solving these two equations simultaneously, we get $x_1 = x_2 = \mu/(2\mu + 1)$. Thus, the optimal solution of the penalty problem can be made arbitrarily close to the solution of the original problem by choosing $\mu$ sufficiently large.

---

**Figure 9.1 Penalty and auxiliary functions.**
Penalty and Barrier Functions

Geometric Interpretation of Penalty Functions

We now use Example 9.1.2 to illustrate the notion of penalty functions geometrically. Suppose that the constraint $h(x) = 0$ is perturbed so that $h(x) = \varepsilon$; that is, $x_1 + x_2 - 1 = \varepsilon$. Thus, we get the following problem:

$$v(\varepsilon) = \text{Minimize } x_1^2 + x_2^2$$

subject to $x_1 + x_2 - 1 = \varepsilon$.

Substituting $x_2 = 1 + \varepsilon - x_1$ into the objective function, the problem reduces to minimizing $x_1^2 + (1 + \varepsilon - x_1)^2$. The optimum occurs where the derivative vanishes, giving $2x_1 - 2(1 + \varepsilon - x_1) = 0$. Therefore, for any given $\varepsilon$, the optimal solution to the above problem is given by $x_1 = x_2 = (1 + \varepsilon)/2$ and has objective value $v(\varepsilon) = (1 + \varepsilon)^2/2$. In addition, for any given $\varepsilon$, the supremum of $x_1^2 + x_2^2$ subject to $x_1 + x_2 - 1 = \varepsilon$ is equal to $\infty$. Therefore, given any point $(x_1, x_2)$ in $\mathbb{R}^2$ with $x_1 + x_2 - 1 = \varepsilon$, its objective value lies in the interval $[(1 + \varepsilon)^2/2, \infty]$.

In other words, the objective values of all points $x$ in $\mathbb{R}^2$ that satisfy $h(x) = \varepsilon$ lie between $(1 + \varepsilon)^2/2$ and $\infty$. In particular, the set $\{[h(x), f(x)) : x \in \mathbb{R}^2\}$ is shown in Figure 9.2. The lower envelope of this set is given by the parabola $(1 + h)^2/2 = (1 + \varepsilon)^2/2 = v(\varepsilon)$. For a fixed $\mu > 0$, the penalty problem is to minimize $f(x) + \mu h^2(x)$ subject to $x \in \mathbb{R}^2$. The contour $f + \mu h^2 = k$ is illustrated in the $(h, f)$ space of Figure 9.2 by a dashed parabola. The intersection of this parabola on the $f$-axis is equal to $k$. So if $f + \mu h^2$ is to be minimized, the parabola must be moved downward as much as possible so that it still has at least one point in common with the shaded set, which describes legitimate combinations of $h$ and $f$ values. This process is continued until the parabola becomes tangential to the shaded set, as shown in Figure 9.2. This means that for this given value of $\mu$, the optimal value of the penalty problem is the intercept of the parabola on the $f$-axis. Note that the optimal solution to the penalty problem is slightly infeasible to the original problem, since $h \neq 0$ at the point of tangency. Furthermore, the optimal objective value of the penalty problem is slightly smaller than the optimal primal objective. Also note that as the value of $\mu$ increases, the parabola $f + \mu h^2$ becomes steeper, and the point of tangency approaches the true optimal solution to the original problem.
Nonconvex Problems

In Figure 9.2 we showed that penalty functions can be used to get arbitrarily close to the optimal solution of the convex problem of Example 9.1.2. Figure 9.3 shows a nonconvex case in which the Lagrangian dual approach would fail to produce an optimal solution for the primal problem because of the presence of a duality gap. Since penalty functions use a nonlinear support, as opposed to the linear support used by the dual function, shown in Figure 9.3, penalty functions can dip through the shaded set and get arbitrarily close to an optimal solution for the original problem, provided, of course, that a sufficiently large penalty parameter $\mu$ is used.

Interpretation Via Perturbation Functions

Observe that the function $\nu(\varepsilon)$ defined above and illustrated in Figures 9.2 and 9.3 is precisely the perturbation function defined in Equation (6.9). In fact, for
the problem to minimize $f(x)$ subject to $h_i(x) = 0$ for $i = 1, \ldots, \ell$, we have, denoting $\epsilon = (\epsilon_1, \ldots, \epsilon_\ell)'$ as a vector of perturbations,

$$\min_x \left\{ f(x) + \mu \sum_{i=1}^{\ell} h_i^2(x) \right\} = \min_{(x, \epsilon)} \{ f(x) + \mu \| \epsilon \|^2 : h_i(x) = \epsilon_i \text{ for } i = 1, \ldots, \ell \}$$

$$= \min_{\epsilon} \{ \mu \| \epsilon \|^2 + \min_x \{ f(x) : h_i(x) = \epsilon_i \text{ for } i = 1, \ldots, \ell \} \} \quad (9.2)$$

$$= \min_{\epsilon} \{ \mu \| \epsilon \|^2 + \nu(\epsilon) \}.$$

Thus, we intuitively see that even if $\nu$ is nonconvex, as $\mu$ increases, the net effect of adding the term $\mu \| \epsilon \|^2$ to $\nu(\epsilon)$ is to convexify it; and as $\mu \to \infty$, the minimizing $\epsilon$ in (9.2) approaches zero. This interpretation readily extends to include inequality constraints as well (see Exercise 9.11).

In particular, relating this to Figures 9.2 and 9.3 for the case $\ell = 1$, if $x_\mu$ minimizes (9.2) with $h(x_\mu) = \epsilon_\mu$, assuming that $\nu$ is differentiable, we see that for a given $\mu > 0$, $\nu'(\epsilon_\mu) = -2\mu \epsilon_\mu = -2\mu h(x_\mu)$ at the minimizing solution. Moreover, equating the objective values of the first and last minimization problems in (9.2), we obtain $f(x_\mu) = \nu(\epsilon_\mu)$. Hence, the coordinate $[h(x_\mu), f(x_\mu)]$ lies on the graph of $\nu(\epsilon_\mu)$, coinciding with $[\epsilon_\mu, \nu(\epsilon_\mu)]$ and having the slope of $\nu$ at $\epsilon_\mu$ equal to $-2\mu \epsilon_\mu$. Denoting $f(x_\mu) + \mu h^2(x_\mu) = k_\mu$, we see that the parabolic function of $\epsilon$ given by $f = k_\mu - \mu \epsilon^2$ equals $\nu(\epsilon_\mu)$ in value when $\epsilon = \epsilon_\mu$ and has a slope of $-2\mu \epsilon_\mu$ at this point. Therefore, the solution $[h(x_\mu), f(x_\mu)]$ appears as shown in Figures 9.2 and 9.3. Observe also in Figure 9.3 that there does not exist a supporting hyperplane to the epigraph of $\nu$ at the point $[0, \nu(0)]$, thereby leading to a duality gap for the associated Lagrangian dual as shown in Theorem 6.2.7.

### 9.2 Exterior Penalty Function Methods

In this section we present and prove an important result that justifies using exterior penalty functions as a means for solving constrained problems. We also discuss some computational difficulties associated with penalty functions and present some approaches geared toward overcoming such problems. Consider the following primal and penalty problems.
Primal Problem

Minimize \( f(x) \)
subject to \( g(x) \leq 0 \)
\( h(x) = 0 \)
\( x \in X \),

where \( g \) is a vector function with components \( g_1, \ldots, g_m \) and \( h \) is a vector function with components \( h_1, \ldots, h_t \). Here, \( f, g_1, \ldots, g_m, h_1, \ldots, h_t \) are continuous functions defined on \( \mathbb{R}^n \), and \( X \) is a nonempty set in \( \mathbb{R}^n \). The set \( X \) might typically represent simple constraints that could easily be handled explicitly, such as lower and upper bounds on the variables.

Penalty Problem

Let \( \alpha \) be a continuous function of the form (9.1a) satisfying the properties stated in (9.1b). The basic penalty function approach attempts to find

\[
\sup_{\mu \geq 0} \theta(\mu)
\]

where \( \theta(\mu) = \inf\{f(x) + \mu\alpha(x) \mid x \in X\} \). The main theorem of this section states that

\[
\inf\{f(x) \mid x \in X, g(x) \leq 0, h(x) = 0\} = \sup_{\mu \geq 0} \theta(\mu) = \lim_{\mu \to \infty} \theta(\mu).
\]

From this result it is clear that we can get arbitrarily close to the optimal objective value of the primal problem by computing \( \theta(\mu) \) for a sufficiently large \( \mu \). This result is established in Theorem 9.2.2. First, however, the following lemma is needed.

9.2.1 Lemma

Suppose that \( f, g_1, \ldots, g_m, h_1, \ldots, h_t \) are continuous functions on \( \mathbb{R}^n \), and let \( X \) be a nonempty set in \( \mathbb{R}^n \). Let \( \alpha \) be a continuous function on \( \mathbb{R}^n \) given by (9.1), and suppose that for each \( \mu \), there is an \( x_\mu \in X \) such that \( \theta(\mu) = f(x_\mu) + \mu\alpha(x_\mu) \). Then, the following statements hold true:

1. \( \inf\{f(x) \mid x \in X, g(x) \leq 0, h(x) = 0\} \geq \sup_{\mu \geq 0} \theta(\mu) \), where \( \theta(\mu) = \inf\{f(x) + \mu\alpha(x) \mid x \in X\} \) and where \( g \) is the vector function whose components are \( g_1, \ldots, g_m \) and \( h \) is the vector function whose components are \( h_1, \ldots, h_t \).
2. \( f(x_{\mu}) \) is a nondecreasing function of \( \mu \geq 0 \), \( \theta(\mu) \) is a non-decreasing function of \( \mu \), and \( \alpha(x_{\mu}) \) is a nonincreasing function of \( \mu \).

**Proof**

Consider \( x \in X \) with \( g(x) \leq 0 \) and \( h(x) = 0 \), and note that \( \alpha(x) = 0 \). Let \( \mu \geq 0 \).

\[
f(x) = f(x) + \mu \alpha(x) \geq \inf \{ f(y) + \mu \alpha(y) : y \in X \} = \theta(\mu).
\]

Thus, Statement 1 follows. To establish Statement 2, let \( \lambda < \mu \). By the definition of \( \theta(\lambda) \) and \( \theta(\mu) \), the following two inequalities hold true:

\[
\begin{align*}
f(x_{\mu}) + \lambda \alpha(x_{\mu}) &\geq f(x_{\lambda}) + \lambda \alpha(x_{\lambda}) \quad (9.3a) \\
f(x_{\lambda}) + \mu \alpha(x_{\lambda}) &\geq f(x_{\mu}) + \mu \alpha(x_{\mu}). \quad (9.3b)
\end{align*}
\]

Adding these two inequalities and simplifying, we get

\[
(\mu - \lambda)[\alpha(x_{\lambda}) - \alpha(x_{\mu})] \geq 0.
\]

Since \( \mu > \lambda \), we get \( \alpha(x_{\lambda}) \geq \alpha(x_{\mu}) \). It then follows from (9.3a) that \( f(x_{\mu}) \geq f(x_{\lambda}) \) for \( \lambda \geq 0 \). By adding and subtracting \( \mu \alpha(x_{\mu}) \) to the left-hand side of (9.3a), we get

\[
f(x_{\mu}) + \mu \alpha(x_{\mu}) + (\lambda - \mu)\alpha(x_{\mu}) \geq \theta(\lambda).
\]

Since \( \mu > \lambda \) and \( \alpha(x_{\mu}) \geq 0 \), the above inequality implies that \( \theta(\mu) \geq \theta(\lambda) \). This completes the proof.

**9.2.2 Theorem**

Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m \\
& \quad h_i(x) = 0 \quad \text{for } i = 1, \ldots, \ell \\
& \quad x \in X,
\end{align*}
\]

where \( f, g_1, \ldots, g_m, h_1, \ldots, h_\ell \) are continuous functions on \( R^n \) and \( X \) is a nonempty set in \( R^n \). Suppose that the problem has a feasible solution, and let \( \alpha \) be a continuous function given by (9.1). Furthermore, suppose that for each \( \mu \) there exists a solution \( x_{\mu} \in X \) to the problem to minimize \( f(x) + \mu \alpha(x) \) subject to \( x \in X \), and that \( \{x_{\mu}\} \) is contained in a compact subset of \( X \). Then
\[
\inf\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\} = \sup_{\mu \geq 0} \theta(\mu) = \lim_{\mu \to \infty} \theta(\mu),
\]
where \(\theta(\mu) = \inf\{f(x) + \mu\alpha(x) : x \in X\} = f(x_\mu) + \mu\alpha(x_\mu)\). Furthermore, the limit \(\bar{x}\) of any convergent subsequence of \(\{x_\mu\}\) is an optimal solution to the original problem, and \(\mu\alpha(x_\mu) \to 0\) as \(\mu \to \infty\).

**Proof**

By Part 2 of Lemma 9.2.1, \(\theta(\mu)\) is monotone, so that \(\sup_{\mu \geq 0} \theta(\mu) = \lim_{\mu \to \infty} \theta(\mu)\). We first show that \(\alpha(x_\mu) \to 0\) as \(\mu \to \infty\). Let \(y\) be a feasible point and \(\varepsilon > 0\). Let \(x_1\) be an optimal solution to the problem to minimize \(f(x) + \mu\alpha(x)\) subject to \(x \in X\) for \(\mu = 1\). If \(\mu \geq (1/\varepsilon) |f(y) - f(x_1)| + 2\), then, by Part 2 of Lemma 9.2.1, we must have \(f(x_\mu) \geq f(x_1)\).

We now show that \(\alpha(x_\mu) \leq \varepsilon\). By contradiction, suppose that \(\alpha(x_\mu) > \varepsilon\). Noting Part 1 of Lemma 9.2.1, we get
\[
\inf\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\} \geq \theta(\mu) = f(x_\mu) + \mu\alpha(x_\mu) \geq f(x_1) + \mu\alpha(x_\mu) > f(x_1) + |f(y) - f(x_1)| + 2\varepsilon > f(y).
\]
The above inequality is not possible in view of the feasibility of \(y\). Thus, \(\alpha(x_\mu) \leq \varepsilon\) for all \(\mu \geq (1/\varepsilon) |f(y) - f(x_1)| + 2\). Since \(\varepsilon > 0\) is arbitrary, \(\alpha(x_\mu) \to 0\) as \(\mu \to \infty\). Now let \(\{x_{\mu_k}\}\) be any convergent subsequence of \(\{x_\mu\}\), and let \(\bar{x}\) be its limit. Then
\[
\sup_{\mu \geq 0} \theta(\mu) \geq \theta(\mu_k) = f(x_{\mu_k}) + \mu_k\alpha(x_{\mu_k}) \geq f(x_{\mu_k}).
\]
Since \(x_{\mu_k} \to \bar{x}\) and \(f\) is continuous, the above inequality implies that
\[
\sup_{\mu \geq 0} \theta(\mu) \geq f(\bar{x}). \tag{9.4}
\]
Since \(\alpha(x_\mu) \to 0\) as \(\mu \to \infty\), \(\alpha(\bar{x}) = 0\); that is, \(\bar{x}\) is a feasible solution to the original problem. In view of (9.4) and Part 1 of Lemma 9.2.1, it follows that \(\bar{x}\) is an optimal solution to the original problem and that \(\sup_{\mu \geq 0} \theta(\mu) = f(\bar{x})\). Note that \(\mu\alpha(x_\mu) = \theta(\mu) - f(x_\mu)\). As \(\mu \to \infty\), \(\theta(\mu)\) and \(f(x_\mu)\) both approach \(f(\bar{x})\), and hence \(\mu\alpha(x_\mu)\) approaches zero. This completes the proof.

**Corollary**

If \(\alpha(x_\mu) = 0\) for some \(\mu\), then \(x_\mu\) is an optimal solution to the problem.
Proof

If \( \alpha(x_\mu) = 0 \), then \( x_\mu \) is a feasible solution to the problem. Furthermore, since

\[
\inf\{f(x): g(x) \leq 0, h(x) = 0, x \in X\} \geq \theta(\mu) = f(x_\mu) + \mu \alpha(x_\mu) = f(x_\mu),
\]

it follows immediately that \( x_\mu \) is an optimal solution.

Note the significance of the assumption that \( \{x_\mu\} \) is contained in a compact subset in \( X \). Obviously, this assumption holds true if \( X \) is compact. Without this assumption, it is possible that the optimal objective values of the primal and penalty problems are not equal (see Exercise 9.6). This assumption is not restrictive in most practical cases, since the variables usually lie between finite lower and upper bounds.

From Theorem 9.2.2 it follows that the optimal solution \( x_\mu \) to the problem to minimize \( f(x) + \mu \alpha(x) \) subject to \( x \in X \) can be made arbitrarily close to the feasible region by choosing \( \mu \) large enough. Furthermore, by choosing \( \mu \) large enough, \( f(x_\mu) + \mu \alpha(x_\mu) \) can be made arbitrarily close to the optimal objective value of the original primal problem. As we discuss later in this section, one popular scheme for solving the penalty problem is to solve a sequence of problems of the form:

\[
\text{Minimize} \quad f(x) + \mu \alpha(x) \\
\text{subject to} \quad x \in X
\]

for an increasing sequence of penalty parameters. The optimal points \( \{x_\mu\} \) are generally infeasible, but as seen in the proof of Theorem 9.2.2, as the penalty parameter \( \mu \) is made large, the points generated approach an optimal solution from outside the feasible region. Hence, as mentioned previously, this technique is also referred to as an exterior penalty function method.

KKT Lagrange Multipliers at Optimality

Under certain conditions, we can use the solutions to the sequence of penalty problems to recover the KKT Lagrange multipliers associated with the constraints at optimality. Toward this end, suppose that \( X = \mathbb{R}^n \) for simplicity, and consider the primal problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0, i = 1, \ldots, m, \) and \( h_i(x) = 0, i = 1, \ldots, \ell \). (The following analysis generalizes readily to the case where some inequality and/or equality constraints define \( X \); see Exercise 9.12.) Suppose that the penalty function \( \alpha \) is given by (9.1), where, in addition, \( \phi \) and \( \psi \) are continuously differentiable with \( \phi'(y) \geq 0 \) for all \( y \) and \( \phi'(y) = 0 \) for \( y \leq 0 \). Assuming that the conditions of Theorem 9.2.2 hold true, since \( x_\mu \) solves
the problem to minimize $f(x) + \mu \alpha(x)$, the gradient of the objective function of this penalty problem must vanish at $x_\mu$. This gives

$$\nabla f(x_\mu) + \sum_{i=1}^{m} \mu \phi'[g_i(x_\mu)] \nabla g_i(x_\mu) + \sum_{i=1}^{\ell} \mu \psi'[h_i(x_\mu)] \nabla h_i(x_\mu) = 0$$

for all $\mu$.

Now let $\bar{x}$ be an accumulation point of the generated sequence $\{x_\mu\}$. Without loss of generality, assume that $\{x_\mu\}$ itself converges to $\bar{x}$. Denote $I = \{i : g_i(\bar{x}) = 0\}$ to be the set of inequality constraints that are binding at $\bar{x}$. Since $g_i(\bar{x}) < 0$ for all $i \in I$ by Theorem 9.2.2, we have $g_i(x_\mu) < 0$ for $\mu$ sufficiently large, yielding $\mu \phi'[g_i(x_\mu)] = 0$. Hence, we can write the foregoing identity as

$$\nabla f(x_\mu) + \sum_{i \in I} (u_{\mu})_i \nabla g_i(x_\mu) + \sum_{i=1}^{\ell} (v_{\mu})_i \nabla h_i(x_\mu) = 0$$

(9.5a)

for all $\mu$ large enough,

where $u_{\mu}$ and $v_{\mu}$ are vectors having components

$$(u_{\mu})_i = \mu \phi'[g_i(x_\mu)] \geq 0 \text{ for all } i \in I, \text{ and } (v_{\mu})_i = \mu \psi'[h_i(x_\mu)]$$

(9.5b)

for all $i = 1, ..., \ell$

Let us now assume that $\bar{x}$ is a regular solution as defined in Theorem 4.3.7. Then we know that there exist unique Lagrangian multipliers $\bar{u}_i \geq 0, i \in I$, and $\bar{v}_i, i = 1, ..., \ell$, such that

$$\nabla f(\bar{x}) + \sum_{i \in I} \bar{u}_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} \bar{v}_i \nabla h_i(\bar{x}) = 0.$$

(9.5c)

Since $g$, $h$, $\phi$, and $\psi$ are all continuously differentiable and since $\{x_\mu\} \to \bar{x}$, which is a regular point, we must then have in (9.5) that $(u_{\mu})_i \to \bar{u}_i$ for all $i \in I$, and $(v_{\mu})_i \to \bar{v}_i$ for all $i = 1, ..., \ell$.

Hence, for sufficiently large values of $\mu$, the multipliers given by (9.5b) can be used to estimate the KKT Lagrange multipliers at optimality. For example, if $\alpha$ is the quadratic penalty function given by $\alpha(x) = \sum_{i=1}^{m} \max\{0, g_i(x)\}^2 + \sum_{i=1}^{\ell} h_i^2(x)$, then $\phi(y) = [\max\{0, y\}]^2$, $\phi'(y) = 2 \max\{0, y\}$, $\psi(y) = y^2$, and $\psi'(y) = 2y$. Hence, from (9.5b), we obtain
Penalty and Barrier Functions

\[(u_\mu)_i = 2\mu \max \{0, g_i(x_\mu)\} \quad \text{for all } i \in I\]

\[(v_\mu)_i = 2\mu h_i(x_\mu) \quad \text{for all } i = 1, \ldots, \ell\]  \hspace{1cm} (9.6)

In particular, observe that if \(\tilde{u}_i > 0\) for some \(i \in I\), then \((u_\mu)_i > 0\) for \(\mu\) large enough, which in turn implies from (9.6) that \(g_i(x_\mu) > 0\). This means that \(g_i(x) \leq 0\) is violated all along the trajectory leading to \(\bar{x}\), and in the limit, \(g_i(\bar{x}) = 0\). Hence, if \(\tilde{u}_i > 0\) for all \(i \in I\), \(\tilde{v}_j \neq 0\) for all \(j\), then all the constraints that are binding at \(\bar{x}\) are violated along the trajectory \(\{x_\mu\}\) leading to \(\bar{x}\). This therefore motivates the name exterior penalty function method. For instance, in Example 9.1.2 we have \(x_\mu = [\mu/(2\mu + 1), \mu/(2\mu + 1)]\), \(h(x_\mu) = -1/(2\mu + 1)\); so \(v_\mu = -2\mu/(2\mu + 1)\) from (9.6). Note that as \(\mu \to \infty\), \(v_\mu \to -1\), the optimal value of the Lagrange multiplier for this example.

Computational Difficulties Associated with Penalty Functions

The solution to the penalty problem can be made arbitrarily close to an optimal solution to the original problem by choosing \(\mu\) sufficiently large. However, if we choose a very large \(\mu\) and attempt to solve the penalty problem, we might incur some computational difficulties associated with ill-conditioning. With a large \(\mu\), more emphasis is placed on feasibility, and most procedures for unconstrained optimization will move quickly toward a feasible point. Even though this point may be far from the optimum, premature termination could occur. To illustrate, suppose that during the course of optimization we reached a feasible point with \(\alpha(x) = 0\). Especially in the presence of nonlinear equality constraints, a movement from \(x\) along any direction \(d\) may result in infeasible points or feasible points having large objective values. In both cases, the value of the auxiliary function \(f(x + \lambda d) + \mu \alpha(x + \lambda d)\) is greater than \(f(x) + \mu \alpha(x)\) for noninfinitesimal values of the step size \(\lambda\). This is obvious in the latter case. In the former case, \(\alpha(x + \lambda d) > 0\); and since \(\mu\) is very large, any reduction in \(f(x + \lambda d)\) over \(f(x)\) will usually be offset by the accompanying increase in the term \(\mu \alpha(x + \lambda d)\). Thus, improvement is possible only if the step size \(\lambda\) is very small such that the term \(\mu \alpha(x + \lambda d)\) would be small, despite the fact that \(\mu\) is very large. In this case, an improvement in \(f(x + \lambda d)\) over \(f(x)\) may offset the fact that \(\mu \alpha(x + \lambda d) > 0\). The need for using a very small step size may result in slow convergence and premature termination.

The foregoing intuitive discussion also has a formal theoretical basis. To gain insight into this issue, consider the equality constrained problem of minimizing \(f(x)\), subject to \(h_i(x) = 0\), for \(i = 1, \ldots, \ell\). Let \(F(x) = f(x) + \mu \sum_{i=1}^{\ell} \psi[h_i(x)]\) denote the penalized objective function constructed according to (9.1), where \(\psi\) is assumed to be twice differentiable. Then, denoting by \(V\) and
\( \nabla^2 \) the gradient and the Hessian operators, respectively, for the functions \( F, f, \) and \( h_i, i = 1, \ldots, \ell, \) and denoting the first and second derivatives of \( \psi \) by \( \psi' \) and \( \psi'', \) respectively, we get, assuming twice differentiability, that

\[
\nabla^2 F(x) = \nabla^2 f(x) + \mu \sum_{i=1}^{\ell} \psi'[h_i(x)] \nabla h_i(x)
\]

\[
\nabla^4 F(x) = \left[ \nabla^2 f(x) + \sum_{i=1}^{\ell} \mu \psi'[h_i(x)] \nabla^2 h_i(x) \right] + \mu \sum_{i=1}^{\ell} \psi'' h_i(x) \nabla h_i(x) y'.
\] (9.7)

Observe that if we also had inequality constraints present in the problem, and we had used the penalty function (9.1a) with \( \phi(y) = [\max\{0, y\}]^2, \) for example, then \( \phi'(y) = 2 \max\{0, y\}, \) but \( \phi''(y) \) would have been undefined at \( y = 0. \) Hence, \( \nabla^2 F(x) \) would be undefined at points having active inequality constraints. However, if \( y > 0, \) then \( \phi'' = 2, \) so \( \nabla^2 F(x) \) would be defined at points that violate all the inequality constraints; and in such a case, (9.7) would inherit a similar expression as for the equality constraints.

Now, as we know from Chapter 8, the convergence rate behavior of algorithms used for minimizing \( F \) will be governed by the eigenvalue structure of \( \nabla^2 F. \) To estimate this characteristic, let us examine the eigenstructure of (9.7) as \( \mu \to \infty \) and, under the conditions of Theorem 9.2.2, as \( x = x_\mu \to \bar{x}, \) an optimum to the given problem. Assuming that \( \bar{x} \) is a regular solution, we have, from (9.5), that \( \mu \psi'[h_i(x_\mu)] \to \bar{v}_i, \) the optimal Lagrange multiplier associated with the \( i \)th constraint, for \( i = 1, \ldots, \ell. \) Hence, the term within \([\cdot]\) in (9.7) approaches the Hessian of the Lagrangian function \( L(x) = f(x) + \sum_{i=1}^{\ell} \bar{v}_i h_i(x). \)

The other term in (9.7), however, is strongly tied in with \( \mu \) and is potentially explosive. For example, if \( \psi(y) = y^2 \) as in the popular quadratic penalty function, this term equals \( 2\mu \) times a matrix that approaches \( \sum_{i=1}^{\ell} \nabla h_i(\bar{x}) \nabla h_i(\bar{x}) y', \) a matrix of rank \( \ell. \) It can then be shown (see the Notes and References section) that as \( \mu \to \infty, \) we have \( x \equiv x_\mu \to \bar{x}, \) and \( \nabla^2 F \) has \( \ell \) eigenvalues that approach \( \infty, \) while \( n - \ell \) eigenvalues approach some finite limits. Consequently, we can expect a severely ill-conditioned Hessian matrix for large values of \( \mu. \)

Examining the analysis evolving around (8.18), the steepest descent method under such a severe situation would probably be disastrous. On the other hand, Newton's method or its variants, such as conjugate gradient or quasi-Newton methods [operated as, at least, an \((\ell + 1)\)-step process], would be unaffected by the foregoing eigenvalue structure. Superior \( n \)-step superlinear (or superlinear) convergence rates might then be achievable, as discussed in Sections 8.6 and 8.8.
9.2.3 Example

Consider the problem of Example 9.1.2. The penalized objective function \( F \) is given by \( F(x) = x_1^2 + x_2^2 + \mu(x_1 + x_2 - 1)^2 \). Its Hessian, as in (9.7), is given by

\[
\nabla^2 F(x) = \begin{bmatrix}
2(1 + \mu) & 2\mu \\
2\mu & 2(1 + \mu)
\end{bmatrix}.
\]

The eigenvalues of this matrix, computed readily by means of the equation \( \det(\nabla^2 F(x) - \lambda I) = 0 \), are \( \lambda_1 = 2 \) and \( \lambda_2 = 2(1 + 2\mu) \), with respective eigenvectors \( (1, -1) \) and \( (1, 1) \). Note that \( \lambda_2 \to \infty \) as \( \mu \to \infty \), while \( \lambda_1 = 2 \) is finite; hence, the condition number of \( \nabla^2 F \) approaches \( \infty \) as \( \mu \to \infty \). Figure 9.4 depicts the contours of \( F \) for a particular value of \( \mu \). These contours are elliptical, with their major and minor axes oriented along the eigenvectors (see Appendix A.1), becoming more and more steep along the direction \( (1, 1) \) as \( \mu \) increases. Hence, for a large value of \( \mu \), the steepest descent method would severely zigzag to the optimum unless it is fortunately initialized at a convenient starting solution.

---

Figure 9.4 Ill-conditioning effect of a large \( \mu \) value.
Summary of Penalty Function Methods

As a result of the difficulties associated with large penalty parameters described above, most algorithms using penalty functions employ a sequence of increasing penalty parameters. With each new value of the penalty parameter, an optimization technique is employed, starting with the optimal solution obtained for the parameter value chosen previously. Such an approach is sometimes referred to as a *sequential unconstrained minimization technique* (SUMT).

We present below a summary of penalty function methods to solve the problem for minimizing $f(x)$ subject to $g(x) \leq 0$, $h(x) = 0$, and $x \in X$. The penalty function $\alpha$ used is of the form specified in (9.1). These methods do not impose any restriction on $f$, $g$, and $h$ other than that of continuity. However, they can effectively be used only in those cases where an efficient solution procedure is available to solve the problem specified in Step 1 below.

**Initialization Step**  Let $\varepsilon > 0$ be a termination scalar. Choose an initial point $x_1$, a penalty parameter $\mu_1 > 0$, and a scalar $\beta > 1$. Let $k = 1$, and go to the Main Step.

**Main Step**

1. Starting with $x_k$, solve the following problem:

   Minimize $f(x) + \mu_k \alpha(x)$
   subject to $x \in X$.

   Let $x_{k+1}$ be an optimal solution and go to Step 2.

2. If $\mu_k \alpha(x_{k+1}) < \varepsilon$, stop; otherwise, let $\mu_{k+1} = \beta \mu_k$, replace $k$ by $k + 1$, and go to Step 1.

**9.2.4 Example**

Consider the following problem:

Minimize $(x_1 - 2)^4 + (x_1 - 2x_2)^2$
subject to $x_1^2 - x_2 = 0$
$x \in X = R^2$.

Note that at iteration $k$, for a given penalty parameter $\mu_k$, the problem to be solved for obtaining $x_{\mu_k}$ is, using the quadratic penalty function:

Minimize $(x_1 - 2)^4 + (x_1 - 2x_2)^2 + \mu_k (x_1^2 - x_2)^2$.

Table 9.1 summarizes the computations using the penalty function method, including the Lagrange multiplier estimate obtained via (9.6). The starting point is taken as $x_1 = (2.0, 1.0)$, where the objective function value is 0.0. The initial
value of the penalty parameter is taken as $\mu_1 = 0.1$, and the scalar $\beta$ is taken as 10.0. Note that $f(x_{\mu_k})$ and $\theta(\mu_k)$ are nondecreasing functions and that $\alpha(x_{\mu_k})$ is a nonincreasing function. The procedure could have been stopped after the fourth iteration, where $\alpha(x_{\mu_k}) = 0.000267$. However, to show more clearly that $\mu_k\alpha(x_{\mu_k})$ does converge to zero according to Theorem 9.2.2, one more iteration was carried out. At the point $x' = (0.9461094, 0.8934414)$, the reader can verify that the KKT conditions are satisfied with the Lagrange multiplier equal to 3.3632. Figure 9.5 shows the progress of the algorithm.

9.3 Exact Absolute Value and Augmented Lagrangian Penalty Methods

For the types of penalty functions considered thus far, we have seen that we need to make the penalty parameter infinitely large in a limiting sense to recover an optimal solution. This can cause numerical difficulties and ill-conditioning effects. A natural question to raise, then, is: Can we design a penalty function that is capable of recovering an exact optimal solution for reasonable finite values of the penalty parameter $\mu$ without the need for $\mu$ to approach infinity? We present below two penalty functions that possess this property and are therefore known as exact penalty functions.

Figure 9.5 Penalty function method.
Table 9.1 Summary of Computations for the Penalty Function Method

<table>
<thead>
<tr>
<th>Iteration ( k )</th>
<th>( \mu_k )</th>
<th>( x_{k+1} = x_{\mu_k} )</th>
<th>( f(x_{k+1}) )</th>
<th>( \alpha(x_{\mu_k}) = h^2(x_{\mu_k}) )</th>
<th>( \theta(\mu_k) )</th>
<th>( \mu_k \alpha(x_{\mu_k}) )</th>
<th>( \nu_{\mu_k} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
<td>(1.4539, 0.7608)</td>
<td>0.0935</td>
<td>1.8307</td>
<td>0.2766</td>
<td>0.1831</td>
<td>0.270605</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>(1.1687, 0.7407)</td>
<td>0.5753</td>
<td>0.3908</td>
<td>0.9661</td>
<td>0.3908</td>
<td>1.250319</td>
</tr>
<tr>
<td>3</td>
<td>10.0</td>
<td>(0.9906, 0.8425)</td>
<td>1.5203</td>
<td>0.01926</td>
<td>1.7129</td>
<td>0.1926</td>
<td>2.775767</td>
</tr>
<tr>
<td>4</td>
<td>100.0</td>
<td>(0.9507, 0.8875)</td>
<td>1.8917</td>
<td>0.000267</td>
<td>1.9184</td>
<td>0.0267</td>
<td>3.266096</td>
</tr>
<tr>
<td>5</td>
<td>1000.0</td>
<td>(0.9461094, 0.8934414)</td>
<td>1.9405</td>
<td>0.0000028</td>
<td>1.9433</td>
<td>0.0028</td>
<td>3.363252</td>
</tr>
</tbody>
</table>
The absolute value or $\ell_1$ penalty function is an exact penalty function, which conforms with the typical form of (9.1) with $p = 1$. Namely, given a penalty parameter $\mu > 0$, the penalized objective function in this case for Problem $P$ to minimize $f(x)$ subject to $g_i(x) \leq 0$, $i = 1, \ldots, m$, and $h_i(x) = 0$, $i = 1, \ldots, \ell$, is given by

$$F_E(x) = f(x) + \mu \left[ \sum_{i=1}^{m} \max\{0, g_i(x)\} + \sum_{i=1}^{\ell} |h_i(x)| \right].$$

(9.8)

(For convenience, we suppress the type of constraints $x \in X$ in our discussion; the analysis readily extends to include such constraints.) The following result shows that under suitable convexity assumptions (and a constraint qualification), there does exist a finite value of $\mu$ that will recover an optimum solution to $P$ via the minimization of $F_E$. Alternatively, it can be shown that if $\bar{x}$ satisfies the second-order sufficiency conditions for a local minimum of $P$ as stated in Theorem 4.4.2, then for $\mu$ at least as large as in Theorem 9.3.1, $\bar{x}$ will also be a local minimum of $F_E$ (see Exercise 9.13).

**9.3.1 Theorem**

Consider the following Problem $P$:

Minimize $f(x)$

subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$

$h_i(x) = 0$ for $i = 1, \ldots, \ell$.

Let $\bar{x}$ be a KKT point with Lagrangian multipliers $\bar{u}_i$, $i \in I$, and $\bar{v}_i$, $i = 1, \ldots, \ell$ associated with the inequality and the equality constraints, respectively, where $I = \{i \in \{1, \ldots, m\} : g_i(\bar{x}) = 0\}$ is the index set of binding or active inequality constraints. Furthermore, suppose that $f$ and $g_i$, $i \in I$, are convex functions and that $h_i$, $i = 1, \ldots, \ell$ are affine functions. Then, for $\mu \geq \max\{\bar{u}_i, i \in I, |\bar{v}_i|, i = 1, \ldots, \ell\}$, $\bar{x}$ also minimizes the exact $\ell_1$ penalized objective function $F_E$ defined by (9.8).

**Proof**

Since $\bar{x}$ is a KKT point for Problem $P$, it is feasible to $P$ and satisfies

$$\nabla f(\bar{x}) + \sum_{i \in I} \bar{u}_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} \bar{v}_i \nabla h_i(\bar{x}) = 0, \quad \bar{u}_i \geq 0 \text{ for } i \in I.$$

(9.9)

(Moreover, by Theorem 4.3.8, $\bar{x}$ solves $P$.) Now consider the problem of minimizing $F_E(x)$ over $x \in \mathbb{R}^n$. This can equivalently be restated as follows, for any $\mu \geq 0$:
Minimize \( f(x) + \mu \left[ \sum_{i=1}^{m} y_i + \sum_{i=1}^{\ell} z_i \right] \) \tag{9.10a}

subject to \( y_i \geq g_i(x) \) and \( y_i \geq 0 \) for \( i = 1, \ldots, m \) \tag{9.10b}
\( z_i \geq h_i(x) \) and \( z_i \geq -h_i(x) \) for \( i = 1, \ldots, \ell. \) \tag{9.10c}

The equivalence follows easily by observing that for any given \( x \in \mathbb{R}^n \), the maximum value of the objective function in (9.10a), subject to (9.10b) and (9.10c), is realized by taking \( y_i = \max\{0, g_i(x)\} \) for \( i = 1, \ldots, m \) and \( z_i = |h_i(x)| \) for \( i = 1, \ldots, \ell \). In particular, given \( \overline{x} \), define \( \overline{y}_i = \max\{0, g_i(\overline{x})\} \) for \( i = 1, \ldots, m \), and \( \overline{z}_i = |h_i(\overline{x})| = 0 \) for \( i = 1, \ldots, \ell \).

Note that of the inequalities \( y_i \geq g_i(x), i = 1, \ldots, m \), only those corresponding to \( i \in \mathcal{I} \) are binding, while all the other inequalities in (9.10) are binding at \( (\overline{x}, \overline{y}, \overline{z}) \). Hence, for \( (\overline{x}, \overline{y}, \overline{z}) \) to be a KKT point for (9.10), we must find Lagrangian multipliers \( u^+_i, u^-_i, i = 1, \ldots, m \), and \( v^+_i, v^-_i, i = 1, \ldots, \ell \), associated with the respective pairs of constraints in (9.10b) and (9.10c) such that

\[
\nabla f(\overline{x}) + \sum_{i \in \mathcal{I}} u^+_i \nabla g_i(\overline{x}) + \sum_{i=1}^{\ell} (v^+_i - v^-_i) \nabla h_i(\overline{x}) = 0
\]

\[
\mu - u^+_i - u^-_i = 0 \quad \text{for} \quad i = 1, \ldots, m
\]

\[
\mu - v^+_i - v^-_i = 0 \quad \text{for} \quad i = 1, \ldots, \ell
\]

\[
(u^+_i, u^-_i) \geq 0 \quad \text{for} \quad i = 1, \ldots, m
\]

\[
(v^+_i, v^-_i) \geq 0 \quad \text{for} \quad i = 1, \ldots, \ell
\]

\[
u^+_i = 0 \quad \text{for} \quad i \notin \mathcal{I}.
\]

Given that \( \mu \geq \max\{\overline{u}_i, i \in \mathcal{I}, \overline{v}_i, i = 1, \ldots, \ell\} \), we then have, using (9.9), that \( u^+_i = \overline{u}_i \) for all \( i \in \mathcal{I}, u^+_i = 0 \) for all \( i \notin \mathcal{I}, u^-_i = \mu - u^+_i \) for all \( i = 1, \ldots, m \), and \( v^+_i = (\mu + \overline{v}_i)/2 \) and \( v^-_i = (\mu - \overline{v}_i)/2 \) for \( i = 1, \ldots, \ell \) satisfy the foregoing KKT conditions. By Theorem 4.3.8 and the convexity assumptions stated, it follows that \( (\overline{x}, \overline{y}, \overline{z}) \) solves (9.10), so \( \overline{x} \) minimizes \( F_E \). This completes the proof.

9.3.2 Example
Consider the problem of Example 9.1.2. The Lagrangian multiplier associated with the equality constraints in the KKT conditions at the optimum \( \overline{x} = (1/2, \overline{y}_i, \overline{z}_i) \) satisfies

\[
\nabla f(\overline{x}) + \sum_{i \in \mathcal{I}} u^+_i \nabla g_i(\overline{x}) + \sum_{i=1}^{\ell} (v^+_i - v^-_i) \nabla h_i(\overline{x}) = 0
\]

\[
\mu - u^+_i - u^-_i = 0 \quad \text{for} \quad i = 1, \ldots, m
\]

\[
\mu - v^+_i - v^-_i = 0 \quad \text{for} \quad i = 1, \ldots, \ell
\]

\[
(u^+_i, u^-_i) \geq 0 \quad \text{for} \quad i = 1, \ldots, m
\]

\[
(v^+_i, v^-_i) \geq 0 \quad \text{for} \quad i = 1, \ldots, \ell
\]

\[
u^+_i = 0 \quad \text{for} \quad i \notin \mathcal{I}.
\]
1/2 is \( \bar{v} = -2x_1 = -2x_2 = -1 \). The function \( F_E \) defined by (9.8) for a given \( \mu \geq 0 \) is \( F_E(x) = (x_1^2 + x_2^2) + \mu|x_1 + x_2 - 1| \). When \( \mu = 0 \), this is minimized at \((0, 0)\). For \( \mu > 0 \), minimizing \( F_E(x) \) is equivalent to minimizing \( (x_1 + x_2 + \mu) \), subject to \( z \geq (x_1 + x_2 - 1) \) and \( z \geq (-x_1 - x_2 + 1) \). The KKT conditions for the latter problem require that \( 2x_1 + (v^+ - v^-) = 0 \), \( 2x_2 + (v^+ - v^-) = 0 \), \( \mu = v^+ + v^- \), \( v^+[z - x_1 - x_2 + 1] = v^-[z + x_1 + x_2 - 1] = 0 \); moreover, optimality dictates that \( z = |x_1 + x_2 - 1| \). Now if \((x_1 + x_2) < 1\), then we must have \( z = -x_1 - x_2 + 1 \) and \( v^+ = 0 \) and hence \( v^- = \mu, x_1 = \mu/2, \text{and } x_2 = \mu/2 \). This is a KKT point, provided that \( 0 \leq \mu < 1 \). On the other hand, if \((x_1 + x_2) = 1\), then \( z = 0 \), \( x_1 = x_2 = 1/2 = (v^- - v^+)/2 \) and therefore \( v^+ = (\mu - 1)/2 \) and \( v^- = (\mu + 1)/2 \). This is a KKT point, provided that \( \mu \geq 1 \). However, if \((x_1 + x_2) > 1\), we get \( x_1 = x_2 = -v^+/2 \), while \( v^+ = \mu \). Hence, this means that \((x_1 + x_2) = -\mu > 1\), a contradiction to \( \mu \geq 0 \). Consequently, as \( \mu \) increases from 0, the minimum of \( F_E \) occurs at \((\mu/2, \mu/2)\) until \( \mu \) reaches the value 1, after which it remains at \((1/2, 1/2)\), the optimum to the original problem.

**Geometric Interpretation for the Absolute Value Penalty Function**

The absolute value \( (\ell_1) \) exact penalty function can be given a geometric interpretation similar in spirit to that illustrated in Figure 9.2. Let the perturbation function \( v(\varepsilon) \) be as illustrated therein. However, in the present case, we are interested in minimizing \( f(x) + \mu|h(x)| \) subject to \( x \in \mathbb{R}^2 \). For this we wish to find the smallest value of \( k \) so that the contour \( f + \mu|\bar{h}| = k \) maintains contact with the epigraph of \( v \) in the \((h, f)\) space. This is illustrated in Figure 9.6. Observe that under the condition of Theorem 9.3.1, \((\bar{x}, \bar{u}, \bar{v})\) is a saddle point by Theorem 6.2.6, so by Theorem 6.2.7, we have \( v(y) \geq v(0) - (u^\ell, \bar{v}^\ell)y \) for all \( y \in \mathbb{R}^{m+\ell} \). In Example 9.3.2 this translates to the assertion that the hyperplane \( f = v(0) - \bar{v}h = v(0) + h \) supports the epigraph of \( v \) from below at \([0, v(0)]\). Hence, as seen from Figure 9.6, for \( \mu = 1 \) (or greater than 1), minimizing \( F_E \) recovers the optimum to the original problem.

Although we have overcome the necessity of having to increase the penalty parameter \( \mu \) to infinity to recover an optimum solution by using the \( \ell_1 \) penalty function, the admissible value of \( \mu \) as prescribed by Theorem 9.3.1 that accomplishes this is as yet unspecified. As a result, we again need to examine a sequence of increasing \( \mu \) values until, say, a KKT solution is obtained. Again, if
Figure 9.6 Geometric interpretation of the absolute penalty function.

\( \mu \) is too small, the penalty problem might be unbounded; and if \( \mu \) is too large, ill-conditioning occurs. Moreover, a primary difference here is that we need to deal with a nondifferentiable objective function for minimizing \( F_E \), which, as discussed in Section 8.9, does not enjoy solution procedures that are as efficient as for the differentiable case. However, as will be seen in Chapter 10 \( \ell_1 \) penalty functions serve a very useful purpose as *merit functions*, which measure sufficient acceptable levels of descent to ensure convergence in other algorithmic approaches (e.g., successive quadratic programming methods) rather than playing a direct role in the direction finding process itself.

**Augmented Lagrangian Penalty Functions**

Motivated by our discussion thus far, it is natural to raise the question whether we can design a penalty function that not only recovers an exact optimum for finite penalty parameter values but also enjoys the property of being differentiable. The *augmented Lagrangian penalty function* (ALAG), also known as the *multiplier penalty function*, is one such exact penalty function.

For simplicity, let us begin by discussing the case of problems having only equality constraints, for which augmented Lagrangians were introduced, and then extend the discussion to include inequality constraints. Toward this end, consider Problem P of minimizing \( f(x) \) subject to \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \).

We have seen that if we employ the quadratic penalty function problem to minimize \( f(x) + \mu \sum_{i=1}^{\ell} h_i^2(x) \), we typically need to let \( \mu \to \infty \) to obtain a constrained optimum for P. We might then be curious whether if we were to shift the origin of the penalty term to \( \theta = (\theta_i, i = 1, \ldots, \ell) \) and consider the
penalized objective function $f(x) + \mu \sum_{i=1}^{\ell} [h_i(x) - \theta_i]^2$ with respect to the problem in which the constraint right-hand sides are perturbed to $\theta$ from 0, it could become possible to obtain a constrained minimum to the original problem without letting $\mu \to \infty$. In expanded form, the latter objective function is $f(x) - \sum_{i=1}^{\ell} 2\mu \theta_i h_i(x) + \mu \sum_{i=1}^{\ell} h_i^2(x) + \mu \sum_{i=1}^{\ell} \theta_i^2$. Denoting $v_i = -2\mu \theta_i$ for $i = 1, \ldots, m$ and dropping the final constant term, this can be rewritten as

$$F_{ALAG}(x, v) = f(x) + \sum_{i=1}^{\ell} v_i h_i(x) + \mu \sum_{i=1}^{\ell} h_i^2(x). \quad (9.11)$$

Now observe that if $(\bar{x}, \bar{v})$ is a primal-dual KKT solution for $P$, then indeed at $v = \bar{v}$,

$$\nabla_x F_{ALAG}(\bar{x}, \bar{v}) = \left[ \nabla f(\bar{x}) + \sum_{i=1}^{\ell} \bar{v}_i \nabla h_i(\bar{x}) \right] + 2\mu \sum_{i=1}^{\ell} h_i(\bar{x}) \nabla h_i(\bar{x}) = 0 \quad (9.12)$$

for all values of $\mu$; whereas this was not necessarily the case with the quadratic penalty function, unless $\nabla f(\bar{x})$ was itself zero. Hence, whereas we needed to take $\mu \to \infty$ to recover $\bar{x}$ in a limiting sense using the quadratic penalty function, it is conceivable that we only need to make $\mu$ large enough (under suitable regularity conditions as enunciated below) for the critical point $\bar{x}$ of $F_{ALAG}(\cdot, \bar{v})$ to turn out to be its (local) minimizer. In this respect, the last term in (9.11) turns out to be a local convexifier of the overall function.

Observe that the function (9.11) is the ordinary Lagrangian function augmented by the quadratic penalty term, hence the name augmented Lagrangian penalty function. Accordingly, (9.11) can be viewed as the usual quadratic penalty function with respect to the following problem that is equivalent to $P$:

$$\text{Minimize } \left\{ f(x) + \sum_{i=1}^{\ell} v_i h_i(x) : h_i(x) = 0 \text{ for } i = 1, \ldots, \ell \right\}. \quad (9.13)$$

Alternatively, (9.11) can be viewed as a Lagrangian function for the following problem that is also equivalent to $P$:

$$\text{Minimize } \left\{ f(x) + \mu \sum_{i=1}^{\ell} h_i^2(x) : h_i(x) = 0 \text{ for } i = 1, \ldots, \ell \right\}. \quad (9.14)$$

Since, from this viewpoint, (9.11) corresponds to the inclusion of a "multiplier-based term" in the quadratic penalty objective function, it is also sometimes called a multiplier penalty function. As we shall see shortly, these viewpoints lead to a rich theory and algorithmic felicity that is not present in either the pure quadratic penalty function approach or the pure Lagrangian duality–based approach.

The following result provides the basis by virtue of which the ALAG penalty function can be classified as an exact penalty function.
9.3.3 Theorem

Consider Problem P to minimize $f(x)$ subject to $h_i(x) = 0$ for $i = 1, \ldots, \ell$, and let the KKT solution $(\tilde{x}, \tilde{v})$ satisfy the second-order sufficiency conditions for a local minimum (see Theorem 4.4.2). Then there exists a $\tilde{\mu}$ such that for $\mu \geq \tilde{\mu}$, the ALAG penalty function $F_{\text{ALAG}}(\cdot, \tilde{v})$, defined for some accompanying $v = \tilde{v}$, also achieves a strict local minimum at $\tilde{x}$. In particular, if $f$ is convex and $h_i$, $i = 1, \ldots, \ell$, are affine, then any minimizing solution $\tilde{x}$ for P also minimizes $F_{\text{ALAG}}(\cdot, \tilde{v})$ for all $\mu \geq 0$.

**Proof**

Since $(\tilde{x}, \tilde{v})$ is a KKT solution, we have from (9.12) that $\nabla_x F_{\text{ALAG}}(\tilde{x}, \tilde{v}) = 0$. Furthermore, letting $G(\tilde{x})$ denote the Hessian of $F_{\text{ALAG}}(\cdot, \tilde{v})$ at $x = \tilde{x}$, we have

$$G(\tilde{x}) = \nabla^2 f(\tilde{x}) + \sum_{i=1}^{\ell} \tilde{v}_i \nabla^2 h_i(\tilde{x}) + 2\mu \sum_{i=1}^{\ell} [h_i(\tilde{x}) \nabla^2 h_i(\tilde{x}) + \nabla h_i(\tilde{x}) \nabla h_i(\tilde{x})^t]$$

$$= \nabla^2 L(\tilde{x}) + 2\mu \sum_{i=1}^{\ell} \nabla h_i(\tilde{x}) \nabla h_i(\tilde{x})^t,$$  \hspace{1cm} (9.15)

where $\nabla^2 L(\tilde{x})$ is the Hessian at $x = \tilde{x}$ of the Lagrangian function for P defined for a multiplier vector $\tilde{v}$. From the second-order sufficiency conditions, we know that $\nabla^2 L(\tilde{x})$ is positive definite on the cone $C = \{d \neq 0 : \nabla h_i(\tilde{x})'d = 0 \text{ for } i = 1, \ldots, \ell\}$.

Now, on the contrary, if there does not exist a $\tilde{\mu}$ such that $G(\tilde{x})$ is positive definite for $\mu \geq \tilde{\mu}$, then it must be the case that given any $\mu_k = k$, $k = 1, 2, \ldots$, there exists a $d_k$ with $\|d_k\| = 1$ such that

$$d'_k G(\tilde{x}) d_k = d'_k \nabla^2 L(\tilde{x}) d_k + 2k \sum_{i=1}^{\ell} [\nabla h_i(\tilde{x})'d_k]^2 \leq 0.$$  \hspace{1cm} (9.16)

Since $\|d_k\| = 1$ for all $k$, there exists a convergent subsequence for $\{d_k\}$ with limit point $\tilde{d}$, where $\|\tilde{d}\| = 1$. Over this subsequence, since the first term in (9.16) approaches $\tilde{d}' \nabla^2 L(\tilde{x}) \tilde{d}$, a constant, we must have that $\nabla h_i(\tilde{x})' \tilde{d} = 0$ for all $i = 1, \ldots, \ell$ for (9.16) to hold true for all $k$. Hence, $\tilde{d} \in C$. Moreover, since $d_k' \nabla^2 L(\tilde{x}) d_k \leq 0$ for all $k$ by (9.16), we have $\tilde{d}' \nabla^2 L(\tilde{x}) \tilde{d} \leq 0$. This contradicts the second-order sufficiency conditions. Consequently, $G(\tilde{x})$ is positive definite.
for \( \mu \) exceeding some value \( \bar{\mu} \), so, by Theorem 4.1.4, \( \bar{x} \) is a strict local minimum for \( F_{\text{ALAG}}(\cdot, \bar{\nu}) \).

Finally, suppose that \( f \) is convex, \( h_i, i = 1, \ldots, \ell \) are affine, and \( \bar{x} \) is optimal to \( P \). By Lemma 5.1.4, there exists a Lagrange multiplier vector \( \bar{\nu} \) such that \((\bar{x}, \bar{\nu})\) is a KKT solution. As before, we have \( \nabla_x F_{\text{ALAG}}(\bar{x}, \bar{\nu}) = 0 \), and since \( F_{\text{ALAG}}(\cdot, \bar{\nu}) \) is convex for any \( \mu \geq 0 \), this completes the proof.

We remark here that without the second-order sufficiency conditions of Theorem 9.3.3, there might not exist any finite value of \( \mu \) that will recover an optimum \( \bar{x} \) for Problem \( P \), and it might be that we need to take \( \mu \rightarrow \infty \) for this to occur. The following example from Fletcher [1987] illustrates this point.

**9.3.4 Example**

Consider Problem \( P \) to minimize \( f(x) = x_1^4 + x_1x_2 \), subject to \( x_2 = 0 \). Clearly, \( \bar{x} = (0, 0)' \) is the optimal solution. From the KKT conditions, we also obtain \( \bar{\nu} = 0 \) as the unique Lagrangian multiplier. Since

\[
\nabla^2 L(\bar{x}) = \nabla^2 f(\bar{x}) = \begin{bmatrix} 0 & 1 \\ 1 & 2\mu \end{bmatrix}
\]

is indefinite, the second-order sufficiency condition does not hold at \((\bar{x}, \bar{\nu})\). Now we have \( F_{\text{ALAG}}(\bar{x}, \bar{\nu}) = x_1^4 + x_1x_2 + \mu x_2^2 = F(x) \), say. Note that for any \( \mu > 0 \),

\[
\nabla F(x) = \begin{pmatrix} 4x_1^3 + x_2 \\ x_1 + 2\mu x_2 \end{pmatrix}
\]

vanishes at \( \bar{x} = (0, 0)' \) and at \( \hat{x} = (1/\sqrt{8\mu}, -1/(2\mu\sqrt{8\mu}))' \). Furthermore,

\[
\nabla^2 F(x) = \begin{bmatrix} 12x_1^2 & 1 \\ 1 & 2\mu \end{bmatrix}
\]

We see that \( \nabla^2 F(\bar{x}) \) is indefinite and hence \( \bar{x} \) is not a local minimizer for any \( \mu > 0 \). However, \( \nabla^2 F(\hat{x}) \) is positive definite, and \( \hat{x} \) is in fact the minimizer of \( F \) for all \( \mu > 0 \). Moreover, as \( \mu \rightarrow \infty \), \( \hat{x} \) approaches the constrained minimum to Problem \( P \).

**9.3.5 Example**

Consider the problem of Example 9.1.2. We have seen that \( \bar{x} = (1/2, 1/2)' \), with \( \bar{\nu} = -1 \), is the unique KKT point and optimum for this problem. Furthermore,
\[ \nabla^2 L(\bar{x}) = \nabla^2 f(\bar{x}) \text{ is positive definite, and thus the second-order sufficiency condition holds true at } (\bar{x}, \bar{v}). \] Moreover, from (9.11), \( F_{\text{LAG}}(\bar{x}, \bar{v}) = (x_1^2 + x_2^2) - (x_1 + x_2 - 1) + \mu(x_1 + x_2 - 1)^2 = (x_1 - 1/2)^2 + (x_2 - 1/2)^2 + \mu(x_1 + x_2 - 1)^2 + 1/2, \) which is clearly uniquely minimized at \( \bar{x} = (1/2, 1/2)' \) for all \( \mu \geq 0. \) Hence, both assertions of Theorem 9.3.3 are verified.

**Geometric Interpretation for the Augmented Lagrangian Penalty Function**

The ALAG penalty function can be given a geometric interpretation similar to that illustrated in Figures 9.2 and 9.3. Let \( v(\epsilon) = \min \{ f(x) : h(x) = \epsilon \} \) be the perturbation function as illustrated therein. Assume that \( \bar{x} \) is a regular point and that \( (\bar{x}, \bar{v}) \) satisfies the second-order sufficiency condition for a strict local minimum. Then it follows readily in the spirit of Theorem 6.2.7 that \( \nabla v(0) = -\bar{v} \) (see Exercise 9.17).

Now, consider the case of Figure 9.2, which illustrates Examples 9.1.2 and 9.3.5. For a given \( \mu > 0, \) the augmented Lagrangian penalty problem seeks the minimum of \( f(x) + \bar{v}h(x) + \mu h^2(x) \) over \( x \in \mathbb{R}^2. \) This amounts to finding the smallest value of \( k \) for which the contour \( f + \bar{v}h + \mu h^2 = k \) maintains contact with the epigraph of \( v \) in the \( (h,f) \) space. This contour equation can be rewritten as \( f = -\mu[(\bar{v}/2\mu)^2 + k + (\bar{v}^2/4\mu)] \) and represents a parabola whose axis is shifted to \( h = -\bar{v}^2/2\mu \) relative to that at \( h = 0 \) as in Figure 9.2. Figure 9.7a illustrates this situation. Note that when \( h = 0, \) we have \( f = k \) on this parabola. Moreover, when \( k = v(0), \) which is the optimal objective value for Problem P, the parabola passes through the point \( (0, v(0)) \) in the \( (h,f) \) plane, and the slope of the tangent to the parabola at this point equals \( -\bar{v}. \) This slope therefore coincides with \( v'(0) = -\bar{v}. \) Hence, as shown in Figure 9.7a, for any \( \mu > 0, \) the minimum of the augmented Lagrangian penalty function coincides with the optimum to Problem P.

For the nonconvex case, under the foregoing assumption that assures \( \nabla v(0) = -\bar{v}, \) a similar situation occurs; but in this case, this happens only when \( \mu \) gets sufficiently large, as shown in Figure 9.7b. In contrast, the Lagrangian dual problem leaves a duality gap, and the quadratic penalty function needs to take \( \mu \to \infty \) to recover the optimal solution.

To gain further insight, observe that we can write the minimization of the ALAG penalty function for any \( v \) in terms of the perturbation function as

\[
\min_{x} \left\{ f(x) + v'(x') + \mu \|x'\|^2 \right\} = \min_{(x', \epsilon)} \left\{ f(x) + v'(\epsilon) + \mu \|\epsilon\|^2 : h(x) = \epsilon \right\}
\]

\[
= \min_{\epsilon} \left\{ v(\epsilon) + v'(\epsilon) + \mu \|\epsilon\|^2 \right\}.
\]
Penalty and Barrier Functions

Optimal solution to penalty problem, 

\[ h(X) = \epsilon \]

Note that if we take \( v = \bar{v} \) and define

\[ V(\epsilon) = v(\epsilon) + \bar{v}^T \epsilon + \mu \| \epsilon \|^2, \]

then when \( \mu \) get sufficiently large, \( V \) becomes a strictly convex function in the neighborhood of \( \epsilon = 0 \); and moreover, \( \nabla V(0) = \nabla v(0) + \bar{v} = 0 \). Hence, we obtain a strict local minimum for \( V \) at \( \epsilon = 0 \). Figure 9.7 illustrates this situation.

Schema of an Algorithm Using Augmented Lagrangian Functions: Method of Multipliers

The method of multipliers is an approach for solving nonlinear programming problems by using the augmented Lagrangian penalty function in a manner that combines the algorithmic aspects of both Lagrangian duality methods and penalty function methods. However, this is accomplished while gaining from both these concepts without being impaired by their respective shortcomings. The method adopts a dual ascent step similar to the subgradient optimization scheme for optimizing the Lagrangian dual; but unlike the latter approach, the overall procedure produces both primal and dual solutions. The primal solution is produced via a penalty function minimization; but because of the properties of the ALAG penalty function, this can usually be accomplished without having to make the penalty parameter infinitely large and, hence, having to contend with...
the accompanying ill-conditioning effects. Moreover, we can employ efficient
derivative-based methods in minimizing the penalized objective function.

The fundamental schema of this type of algorithm is as follows. Consider
the problem of minimizing \( f(x) \) subject to the equality constraints \( h_i(x) = 0 \) for
\( i = 1, \ldots, \ell \). (The extension to incorporate inequality constraints is relatively
straightforward and is addressed in the following subsection.) Below we outline
the procedure first and then provide some interpretations, motivations, and
implementation comments. As is typically the case, the augmented Lagrangian
function employed is of the form (9.11), except that each constraint is assigned
its own specific penalty parameter \( \mu_i \), in lieu of a common parameter \( \mu \). Hence,
constraint violations, and consequent penalizations, can be monitored
individually. Accordingly, we replace (9.11) by

\[
F_{\text{ALAG}}(x, \nu) = f(x) + \sum_{i=1}^{\ell} v_i h_i(x) + \sum_{i=1}^{\ell} \mu_i h_i^2(x).
\]

\( (9.18) \)

**Initialization** Select some initial Lagrangian multiplier vector \( \nu = \bar{\nu} \) and
positive values \( \mu_1, \ldots, \mu_\ell \) for the penalty parameters. Let \( x_0 \) be a null vector,
and denote \( \text{VIOL}(x_0) = \infty \), where for any \( x \in R^n \), \( \text{VIOL}(x) = \max \{ |h_i(x)| : i = 1, \ldots, \ell \} \) is a measure of constraint violations. Put \( k = 1 \) and proceed to the
inner loop of the algorithm.

**Inner Loop: Penalty Function Minimization** Solve the unconstrained
problem to minimize \( F_{\text{ALAG}}(x, \nu) \) subject to \( x \in R^n \), and let \( x_k \) denote the
optimal solution obtained. If \( \text{VIOL}(x_k) = 0 \), stop, with \( x_k \) as a KKT point.
[Practically, we could terminate if \( \text{VIOL}(x_k) \) is less than some tolerance \( \varepsilon > 0 \).] Otherwise, if \( \text{VIOL}(x_k) \leq (1/4)\text{VIOL}(x_{k-1}) \), proceed to the outer loop.
On the other hand, if \( \text{VIOL}(x_k) > (1/4)\text{VIOL}(x_{k-1}) \), then for each constraint
\( i = 1, \ldots, \ell \) for which \( |h_i(x)| > (1/4)\text{VIOL}(x_{k-1}) \), replace the corresponding
penalty parameter \( \mu_i \) by \( 10\mu_i \) and repeat this inner loop step.

**Outer Loop: Lagrange Multiplier Update** Replace \( \nu \) by \( \nu_{\text{new}} \), where

\[
(\nu_{\text{new}})_i = \bar{\nu}_i + 2\mu_i h_i(x_k) \quad \text{for } i = 1, \ldots, \ell.
\]

\( (9.19) \)

Increment \( k \) by 1, and return to the inner loop.

The inner loop of the foregoing method is concerned with the
minimization of the augmented Lagrangian penalty function. For this purpose,
we can use \( x_{k-1} \) (for \( k \geq 2 \)) as a starting solution and employ Newton’s method
(with line searches) in case the Hessian is available, or else use a quasi-Newton
method if only gradients are available, or use some conjugate gradient method
Penalty and Barrier Functions

497

for relatively large-scale problems. If \( \text{VIOL}(x_k) = 0 \), then \( x_k \) is feasible, and moreover,

\[
\nabla_x F_{\text{ALAG}}(x_k, \nu) = \nabla f(x_k) + \sum_{i=1}^\ell \nabla h_i(x_k) + \sum_{i=1}^\ell 2\mu_i h_i(x_k) \nabla h_i(x_k) = 0 \quad (9.20)
\]

implies that \( x_k \) is a KKT point. Whenever the revised iterate \( x_k \) of the inner loop does not improve the measure for constraint violations by the factor 1/4, the penalty parameter is increased by a factor of 10. Hence, the outer loop will be visited after a finite number of iterations when the tolerance \( \varepsilon > 0 \) is used in the inner loop, since as in Theorem 9.2.2, we have \( h_i(x_k) \to 0 \) as \( \mu_i \to \infty \) for \( i = 1, ..., \ell \).

Observe that the foregoing argument holds true regardless of the dual multiplier update scheme used in the outer loop, and that it is essentially related to using the standard quadratic penalty function approach on the equivalent problem (9.13). In fact, if we adopt this viewpoint, the estimate of the Lagrange multipliers associated with the constraints in (9.13) is given by \( 2\mu_i h_i(x_k) \) for \( i = 1, ..., \ell \), as in (9.6). Since the relationship between the Lagrange multipliers of the original Problem P and its primal equivalent form (9.13) with \( \nu = \overline{\nu} \) is that the Lagrange multiplier vector for P equals \( \overline{\nu} \) plus the Lagrange multiplier vector for (9.13), Equation (9.19) then gives the corresponding estimate for the Lagrange multipliers associated with the constraints of P.

This observation can be reinforced more directly by the following interpretation. Note that having minimized \( F_{\text{ALAG}}(x, \overline{\nu}) \), we have (9.20) holding true. However, for \( x_k \) and \( \overline{\nu} \) to be a KKT solution, we want \( \nabla_x L(x_k, \overline{\nu}) = 0 \), where \( L(x, \nu) = f(x) + \sum_{i=1}^\ell \nu_i h_i(x) \) is the Lagrangian function for Problem P. Hence, we can choose to revise \( \overline{\nu} \) to \( \overline{\nu}_{\text{new}} \) in a manner such that \( \nabla f(x_k) + \sum_{i=1}^\ell (\overline{\nu}_{\text{new}})_i \nabla h_i(x_k) = 0 \). Superimposing this identity on (9.20) provides the update scheme (9.19).

Hence, from the viewpoint of the problem (9.13), convergence is obtained above in one of two ways. First, we might finitely determine a KKT point, as is frequently the case. Alternatively, viewing the foregoing algorithm as one of applying the standard quadratic penalty function approach, in spirit, to the equivalent sequence of problems of the type (9.13), each having particular estimates of the Lagrangian multipliers in the objective function, convergence is obtained by letting the penalty parameters approach infinity. In the latter case, the inner loop problems become increasingly ill-conditioned, and second-order methods become imperative.

There is an alternative Lagrangian duality-based interpretation of the update scheme (9.19) when \( \mu_i = \mu \) for all \( i = 1, ..., \ell \), which leads to an improved procedure having a better overall rate of convergence. Recall that Problem P is also equivalent to the problem (9.14), where this equivalence now holds true
with respect to both primal and dual solutions. Moreover, the Lagrangian dual function for (9.14) is given by \( \theta(v) = \min_x \{ F_{\text{LAG}}(x, v) \} \), where \( F_{\text{LAG}}(x, v) \) is given by (9.11). Hence, at \( v = \overline{v} \), the inner loop essentially evaluates \( \theta(\overline{v}) \), determining an optimal solution \( x_k \). This yields \( h(x_k) \) as a subgradient of \( \theta \) at \( v = \overline{v} \). Consequently, the update \( \bar{v}_{\text{new}} = \overline{v} + 2\mu h(x_k) \) as characterized by (9.19) is simply a fixed-step-length subgradient direction-based iteration for the dual function.

This raises the issue that if a quadratically convergent Newton scheme, or a superlinearly convergent quasi-Newton method, is used for the inner loop optimization problems, the advantage of using a second-order method is lost if we employ a linearly convergent gradient-based update scheme for the dual problem. As can be surmised, the convergence rate for the overall algorithm is intimately connected with that of the dual updating scheme. Assuming that the problem (9.14) has a local minimum at \( x^* \), that \( x^* \) is a regular point with the (unique) Lagrange multiplier vector being given by \( v^* \), and that the Hessian of the Lagrangian with respect to \( x \) is positive definite at \( (x^*, v^*) \), then it can be shown (see Exercise 9.10) that in a local neighborhood of \( v^* \), the minimizing solution \( x(v) \) that evaluates \( \theta(v) \) when \( x \) is confined to be near \( x^* \) is a continuously differentiable function. Hence, we have \( \theta(v) = F_{\text{LAG}}[x(v), v] \); and so, since \( \nabla_x F_{\text{LAG}}[x(v), v] = 0 \), we have

\[
\nabla \theta(v) = \nabla_v F_{\text{LAG}}[x(v), v] = h[x(v)]. \tag{9.21}
\]

Denoting \( \nabla h(x) \) and \( \nabla x(v) \) to be the Jacobians of \( h \) and \( x \), respectively, we then have

\[
\nabla^2 \theta(v) = \nabla h[x(v)] \nabla x(v). \tag{9.22a}
\]

Differentiating the identity \( \nabla_x F_{\text{LAG}}[x(v), v] = 0 \) with respect to \( v \), we obtain [see Equation (9.12)]

\[
\nabla^2_x F_{\text{LAG}}[x(v), v] \nabla x(v) + \nabla h[x(v)]^f = 0. \]

Solving for \( \nabla x(v) \) (in the neighborhood of \( v^* \)) from this equation and substituting into (9.22a) gives

\[
\nabla^2 \theta(v) = -\nabla h[x(v)] \{ \nabla^2_x F_{\text{LAG}}[x(v), v] \}^{-1} \nabla h[x(v)]^f. \tag{9.22b}
\]

At \( x^* \) and \( v^* \), the matrix \( \nabla^2_x F_{\text{LAG}}[x^*, v^*] = \nabla^2_x L(x^*) + 2\mu \nabla h(x^*)^f \nabla h(x^*) \), and the eigenvalues of this matrix determine the rate of convergence of the gradient-based algorithm that uses the update scheme (9.19). Also, a second-order quasi-Newton type of an update scheme can employ an approximation \( B \) for \( \{ \nabla^2_x F_{\text{LAG}}[x(v), v] \}^{-1} \) and, accordingly, determine \( \bar{v}_{\text{new}} \) as an approximation to \( \overline{v} - [\nabla^2 \theta(v)]^{-1} \nabla \theta(v) \). Using (9.21) and (9.22b), this gives
Penalty and Barrier Functions

\[ \bar{v}_{\text{new}} = \bar{v} + [\nabla h(x) B \nabla h(x)']^{-1} h(x). \]  
(9.23)

It is also interesting to note that as \( \mu \to \infty \), the Hessian in (9.22b) approaches \((1/2) \mu I\), and (9.23) then approaches (9.19). Hence, as the penalty parameter increases, the condition number of the Hessian in the dual problem becomes close to unity, implying a very rapid outer loop convergence, whereas that in the penalty problem of the inner loop becomes increasingly worse.

9.3.6 Example

Consider the problem of Example 9.1.2. Given any \( v \), the inner loop of the method of multipliers evaluates \( \theta(v) = \min_x \{ F_{\text{LAG}}(x, v) \} \), where \( F_{\text{LAG}}(x, v) = x_1^2 + x_2^2 + v(x_1 + x_2 - 1) + \mu(x_1 + x_2 - 1)^2 \). Solving \( \nabla_x F_{\text{LAG}}(x, v) = 0 \) yields \( x_1(v) = x_2(v) = (2\mu - v)/2(1 + 2\mu) \). The outer loop then updates the Lagrange multiplier according to (9.19), which gives \( v_{\text{new}} = v + 2\mu[x_1(v) + x_2(v) - 1] = (v - 2\mu)/(1 + 2\mu) \). Note that as \( \mu \to \infty \), \( v_{\text{new}} \to -1 \), the optimal Lagrange multiplier value.

Hence, suppose that we begin this algorithm with \( \bar{v} = 0 \) and \( \mu = 1 \). The inner loop determines \( x(0) = (1/3, 1/3)' \), with \( \text{VIOL}[x(0)] = 1/3 \), and the outer loop finds \( v_{\text{new}} = -2/3 \). Next, at the second iteration, the inner loop solution is obtained as \( x(-2/3) = (4/9, 4/9)' \) with \( \text{VIOL}[x(-2/3)] = 1/9 > (1/4)\text{VIOL}[x(0)] \). Hence, we increase \( \mu \) to 10 and recompute the revised \( x(-2/3) = (31/63, 31/63)' \) with \( \text{VIOL}[x(-2/3)] = 1/63 \). The outer loop then revises the Lagrange multiplier \( \bar{v} = -2/3 \) to \( \bar{v}_{\text{new}} = -62/63 \). The iterations proceed in this fashion, using the foregoing formulas, until the constraint violation at the inner loop solution is acceptably small.

Extension to Include Inequality Constraints in the ALAG Penalty Function

Consider Problem P to minimize \( f(x) \) subject to the constraints \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) and \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \). The extension of the foregoing theory of augmented Lagrangians and the method of multipliers to this case, which also includes inequality constraints, is readily accomplished by equivalently writing the inequalities as the equations \( g_i(x) + s_i^2 = 0 \) for \( i = 1, \ldots, m \). Now, suppose that \( \bar{x} \) is a KKT point for Problem P with optimal Lagrange multipliers \( \bar{\mu}_i, i = 1, \ldots, m \), and \( \bar{s}_i, i = 1, \ldots, \ell \), associated with the inequality and the equality constraints, respectively, and such that the strict complementary slackness condition holds true: namely, that \( \bar{s}_i g_i(\bar{x}) = 0 \) for all \( i = 1, \ldots, m \), with \( \bar{\mu}_i > 0 \) for each \( i \in I(\bar{x}) = \{ i : g_i(\bar{x}) = 0 \} \). Furthermore, suppose that the second-order
sufficiency condition of Theorem 4.4.2 holds true at \((\bar{x}, \bar{u}, \bar{v})\): namely, that 
\[ \nabla^2 L(\bar{x}) \text{ is positive definite over the cone } C = \{ d \neq 0 : \nabla g_i(\bar{x})'d = 0 \text{ for all } i \in I(\bar{x}), \nabla h_i(\bar{x})'d = 0 \text{ for all } i = 1, \ldots, \ell \}. \] (Note that \(I^0 = \emptyset\) in Theorem 4.4.2 due to strict complementary slackness.) Then it can readily be verified (see Exercise 9.16) that the conditions of Theorem 9.3.3 are satisfied for Problem \(P'\) to minimize \(f(x)\) subject to the equality constraints \(g_i(x) + s_i^2 = 0 \text{ for } i = 1, \ldots, m,\) and \(h_i(x) = 0 \text{ for } i = 1, \ldots, \ell,\) at the solution \((\bar{x}, \bar{s}, \bar{u}, \bar{v})\), where \(\bar{s}_i^2 = -g_i(\bar{x})\) for all \(i = 1, \ldots, m.\) Hence, for \(\mu\) large enough, the solution \((\bar{x}, \bar{s})\), will turn out to be a strict local minimizer for the following ALAG penalty function at \((u, v) = (\bar{u}, \bar{v})\):

\[
f(x) + \sum_{i=1}^{m} \mu_i [g_i(x) + s_i^2] + \sum_{i=1}^{\ell} v_i h_i(x) + \mu \left[ \sum_{i=1}^{m} (g_i(x) + s_i^2)^2 + \sum_{i=1}^{\ell} h_i^2(x) \right]. \tag{9.24} \]

The representation in (9.24) can be simplified into a more familiar form as follows. For a given penalty parameter \(\mu > 0\), let \(\theta(u,v)\) represent the minimum of (9.24) over \((x, s)\) for any given set of Lagrange multipliers \((u, v)\). Now, let us rewrite (9.24) more conveniently as follows:

\[
f(x) + \mu \sum_{i=1}^{m} \left[ g_i(x) + s_i^2 + \frac{u_i}{2\mu} \right]^2 - \sum_{i=1}^{m} \frac{u_i^2}{4\mu} + \sum_{i=1}^{\ell} v_i h_i(x) + \mu \sum_{i=1}^{\ell} h_i^2(x). \tag{9.25} \]

Hence, in computing \(\theta(u,v)\), we can minimize (9.25) over \((x, s)\) by first minimizing \([g_i(x) + s_i^2 + (u_i/2\mu)]\) over \(s_i\) in terms of \(x\) for each \(i = 1, \ldots, m,\) and then minimizing the resulting expression over \(x \in \mathbb{R}^n.\) The former task is easily accomplished by letting \(s_i^2 = -[g_i(x) + (u_i/2\mu)]\) if this is nonnegative and zero otherwise. Hence, we obtain

\[
\theta(u,v) = \min_x \left\{ f(x) + \mu \sum_{i=1}^{m} \max \left\{ g_i(x) + \frac{u_i}{2\mu}, 0 \right\} - \sum_{i=1}^{m} \frac{u_i^2}{4\mu} \right. \\
\left. + \sum_{i=1}^{\ell} v_i h_i(x) + \mu \sum_{i=1}^{\ell} h_i^2(x) \right\} \tag{9.26} \]

Similar to (9.11), the function \(F_{\text{ALAG}}(x,u,v)\) is sometimes referred to as the **ALAG penalty function** itself in the presence of both inequality and equality constraints. In particular, in the context of the method of multipliers, the inner loop evaluates \(\theta(u,v)\), measures the constraint violations, and revises the
penalty parameter(s) exactly as earlier. If \( x_k \) minimizes (9.26), then the subgradient component of \( \theta(u, v) \) corresponding to \( u_i \) at \((u, v) = (\tilde{u}, \tilde{v})\) is given by
\[
2\mu \max\{g_i(x_k) + (\tilde{u}_i/2\mu), 0\}(1/2\mu) = (-\tilde{u}_i/2\mu) + \max\{g_i(x_k) + (\tilde{u}_i/2\mu), 0\}.
\]
Adopting the fixed step length of \( 2\mu \) along this subgradient direction as for the equality constrained case revises \( u_i \) to \((\tilde{u}_{\text{new}})_i = \tilde{u}_i + 2\mu [-(\tilde{u}_i/2\mu) + \max\{g_i(x_k) + (\tilde{u}_i/2\mu), 0\}] \). Simplifying, this gives
\[
(\tilde{u}_{\text{new}})_i = \tilde{u}_i + \max\{2\mu g_i(x_k), -\tilde{u}_i\} \quad \text{for} \ i = 1, \ldots, m. \tag{9.27}
\]
Alternatively, we can adopt an approximate second-order update scheme (or gradient deflection scheme) as for the case of equality constraints.

### 9.4 Barrier Function Methods

Similar to penalty functions, barrier functions are also used to transform a constrained problem into an unconstrained problem or into a sequence of unconstrained problems. These functions set a barrier against leaving the feasible region. If the optimal solution occurs at the boundary of the feasible region, the procedure moves from the interior to the boundary. The primal and barrier problems are formulated below.

**Primal Problem**

Minimize \( f(x) \)
subject to \( g(x) \leq 0 \)
\( x \in X, \)

where \( g \) is a vector function whose components are \( g_1, \ldots, g_m \). Here \( f, g_1, \ldots, g_m \) are continuous functions defined on \( \mathbb{R}^n \), and \( X \) is a nonempty set in \( \mathbb{R}^n \). Note that equality constraints, if present, are accommodated within the set \( X \). Alternatively, in the case of affine equality constraints, we can possibly eliminate them after solving for some variables in terms of the others, thereby reducing the dimension of the problem. The reason why this treatment is necessary is that barrier function methods require the set \( \{x : g(x) < 0\} \) to be nonempty, which would obviously not be possible if the equality constraints \( h(x) = 0 \) were accommodated within the set of inequalities as \( h(x) \leq 0 \) and \( h(x) \geq 0 \).

**Barrier Problem**

Find:
\[
\inf \theta(\mu) \quad \text{subject to} \ \mu > 0,
\]

where \( \theta(\mu) = \inf\{f(x) + \mu B(x) : g(x) < 0, x \in X\} \). Here \( B \) is a barrier function that is nonnegative and continuous over the region \( \{x : g(x) < 0\} \) and approaches
as the boundary of the region \( \{ x : g(x) \leq 0 \} \) is approached from the interior. More specifically, the barrier function \( B \) is defined by

\[
B(x) = \sum_{i=1}^{m} \phi[g_i(x)],
\]

(9.28a)

where \( \phi \) is a univariate function that is continuous over \( \{ y : y < 0 \} \) and satisfies

\[
\phi(y) \geq 0 \text{ if } y < 0 \quad \text{and} \quad \lim_{y \to 0^-} \phi(y) = \infty.
\]

(9.28b)

For example, a typical barrier function might be of the form

\[
B(x) = \sum_{i=1}^{m} \frac{-1}{g_i(x)} \quad \text{or} \quad B(x) = - \sum_{i=1}^{m} \ln[\min\{1, -g_i(x)\}].
\]

(9.29a)

Note that the second barrier function in (9.29a) is not differentiable because of the term \( \min\{1, -g_i(x)\} \). Actually, since the property (9.28b) for \( \phi \) is essential only in a neighborhood of \( y = 0 \), it can be shown that the following popular barrier function, known as Frisch's logarithmic barrier function,

\[
B(x) = - \sum_{i=1}^{m} \ln[-g_i(x)],
\]

(9.29b)

also admits convergence in the sense of Theorem 9.3.4.

We refer to the function \( f(x) + \mu B(x) \) as the auxiliary function. Ideally, we would like the function \( B \) to take the value zero in the region \( \{ x : g(x) < 0 \} \) and the value \( \infty \) on its boundary. This would guarantee that we would not leave the region \( \{ x : g(x) \leq 0 \} \), provided that the minimization problem started at an interior point. However, this discontinuity poses serious difficulties for any computational procedure. Therefore, this ideal construction of \( B \) is replaced by the more realistic requirement that \( B \) is nonnegative and continuous over the region \( \{ x : g(x) < 0 \} \) and that it approaches infinity as the boundary is approached from the interior. Note that \( \mu B \) approaches the ideal barrier function described above as \( \mu \) approaches zero. Given \( \mu > 0 \), evaluating \( \theta(\mu) = \inf\{f(x) + \mu B(x) : g(x) < 0, x \in X\} \) seems no simpler than solving the original problem because of the presence of the constraint \( g(x) < 0 \). However, as a result of the structure of \( B \), if we start the optimization from a point in the region \( S = \{ x : g(x) < 0 \} \cap X \) and ignore the constraint \( g(x) < 0 \), we will reach an optimal point in \( S \). This results from the fact that as we approach the boundary of \( \{ x : g(x) \leq 0 \} \) from within \( S \), \( B \) approaches infinity, which will prevent us from leaving the set \( S \). This is discussed further in the detailed statement of the barrier function method.

9.4.1 Example

Consider the following problem:
Penalty and Barrier Functions

503

Figure 9.8 Barrier and auxiliary functions.

Minimize \( x \)

subject to \( -x + 1 \leq 0 \).

Note that the optimal solution is \( \bar{x} = 1 \) and that \( f(\bar{x}) = 1 \). Consider the following barrier function:

\[
B(x) = \frac{-1}{-x + 1} \quad \text{for} \ x \neq 1.
\]

Figure 9.8a shows \( \mu B \) for various values of \( \mu > 0 \). Note that as \( \mu \) approaches zero, \( \mu B \) approaches a function that has value zero over \( x > 1 \) and infinity for \( x = 1 \). Figure 9.8b shows the auxiliary function \( f(x) + \mu B(x) = x + [\mu/(x - 1)] \). The dashed functions in Figure 9.8 correspond to the region \( \{x : g(x) > 0\} \) and do not affect the computational process.

Note that for any given \( \mu > 0 \), the barrier problem is to minimize \( x + \mu/(x - 1) \) over the region \( x > 1 \). The function \( x + \mu/(x - 1) \) is convex over \( x > 1 \). Hence, if any of the techniques of Chapter 8 are used to minimize \( x + \mu/(x - 1) \), starting with an interior point \( x > 1 \), we would obtain the optimal point \( x_\mu = 1 + \sqrt{\mu} \). Note that \( f(x_\mu) + \mu B(x_\mu) = 1 + 2\sqrt{\mu} \). Obviously, as \( \mu \rightarrow 0 \), \( x_\mu \rightarrow \bar{x} \) and \( f(x_\mu) + \mu B(x_\mu) \rightarrow f(\bar{x}) \).

We now show the validity of using barrier functions for solving constrained problems by converting them into a single unconstrained problem or into a sequence of unconstrained problems. This is done in Theorem 9.4.3, but first the following lemma is needed.
9.4.2 Lemma

Let \( f, g_1, \ldots, g_m \) be continuous functions on \( \mathbb{R}^n \), and let \( X \) be a nonempty closed set in \( \mathbb{R}^n \). Suppose that the set \( \{ x \in X : g(x) < 0 \} \) is not empty and that \( B \) is a barrier function of the form (9.28) and that it is continuous on \( \{ x : g(x) < 0 \} \). Furthermore, suppose that for any given \( \mu > 0 \), if \( \{ x_k \} \) in \( X \) satisfies \( g(x_k) < 0 \) and \( f(x_k) + \mu B(x_k) \to \theta(\mu) \), then \( \{ x_k \} \) has a convergent subsequence.* Then:

1. For each \( \mu > 0 \), there exists an \( x_\mu \in X \) with \( g(x_\mu) < 0 \) such that
   \[
   \theta(\mu) = f(x_\mu) + \mu B(x_\mu) = \inf\{ f(x) + \mu B(x) : g(x) < 0, x \in X \}. \]

2. \( \inf\{ f(x) : g(x) < 0, x \in X \} \leq \inf\{ \theta(\mu) : \mu > 0 \} \).

3. For \( \mu > 0 \), \( f(x_\mu) \) and \( \theta(\mu) \) are nondecreasing functions of \( \mu \), and \( B(x_\mu) \) is a nonincreasing function of \( \mu \).

**Proof**

Fix \( \mu > 0 \). By the definition of \( \theta \), there exists a sequence \( \{ x_k \} \) with \( x_k \in X \) and \( g(x_k) < 0 \) such that \( f(x_k) + \mu B(x_k) \to \theta(\mu) \). By assumption, \( \{ x_k \} \) has a convergent subsequence \( \{ x_k \} \) with limit \( x_\mu \) in \( X \). By the continuity of \( g \), \( g(x_\mu) \leq 0 \). We show that \( g(x_\mu) < 0 \). If not, then \( g_i(x_\mu) = 0 \) for some \( i \); and since the barrier function \( B \) satisfies (9.28), for \( k \in \mathbb{R} \), \( B(x_k) \to \infty \). Thus, \( \theta(\mu) = \infty \), which is impossible, since \( \{ x : x \in X, g(x) < 0 \} \) is assumed not empty. Therefore, \( \theta(\mu) = f(x_\mu) + \mu B(x_\mu) \), where \( x_\mu \in X \) and \( g(x_\mu) < 0 \), so that Part 1 holds true. Now, since \( B(x) \geq 0 \) if \( g(x) < 0 \), then for \( \mu \geq 0 \), we have

\[
\theta(\mu) = \inf\{ f(x) + \mu B(x) : g(x) < 0, x \in X \} \\
\geq \inf\{ f(x) : g(x) < 0, x \in X \} \\
\geq \inf\{ f(x) : g(x) \leq 0, x \in X \}.
\]

Since the above inequality holds true for each \( \mu \geq 0 \), Part 2 follows. To show Part 3, let \( \mu > \lambda > 0 \). Since \( B(x) \geq 0 \) if \( g(x) < 0 \), then \( f(x) + \mu B(x) \geq f(x) + \lambda B(x) \) for each \( x \in X \) with \( g(x) < 0 \). Thus, \( \theta(\mu) \geq \theta(\lambda) \). Noting Part 1, there exist \( x_\mu \) and \( x_\lambda \) such that

\[
f(x_\mu) + \mu B(x_\mu) \leq f(x_\lambda) + \mu B(x_\lambda) \tag{9.30}
\]

\[
f(x_\lambda) + \lambda B(x_\lambda) \leq f(x_\mu) + \lambda B(x_\mu) \tag{9.31}
\]

* This assumption holds true if \( \{ x \in X : g(x) \leq 0 \} \) is compact.
Adding (9.30) and (9.31) and rearranging, we get 
\((\mu - \lambda)[B(x_\mu) - B(x_\lambda)] \leq 0\).
Since \(\mu - \lambda > 0\), then \(B(x_\mu) \leq B(x_\lambda)\). Substituting this in (9.31), it follows that 
f(\(x_\lambda\)) \leq f(x_\mu). Thus Part 3 holds true, and the proof is complete.

From Lemma 9.4.2, \(\theta\) is a nondecreasing function of \(\mu\) so that 
\(\inf_{\mu > 0} \theta(\mu) = \lim_{\mu \to 0^+} \theta(\mu)\). Theorem 9.4.3 shows that the optimal solution to
the primal problem is indeed equal to \(\lim_{\mu \to 0^+} \theta(\mu)\), so that it could be solved
by a single problem of the form to minimize \(f(x) + \mu B(x)\) subject to \(x \in X\),
where \(\mu\) is sufficiently small, or it can be solved through a sequence of problems
of the above form with decreasing values of \(\mu\).

9.4.3 Theorem

Let \(f: \mathbb{R}^n \to \mathbb{R}\) and \(g: \mathbb{R}^n \to \mathbb{R}^m\) be continuous functions, and let \(X\) be a
nonempty closed set in \(\mathbb{R}^n\). Suppose that the set \(\{x \in X: g(x) < 0\}\) is not empty.
Furthermore, suppose that the primal problem to minimize \(f(x)\) subject to \(g(x) \leq 0, x \in X\) has an optimal solution \(\bar{x}\) with the following property. Given any
neighborhood \(N\) around \(\bar{x}\), there exists an \(x \in X \cap N\) such that \(g(x) < 0\). Then
\[
\min\{f(x) : g(x) \leq 0, x \in X\} = \lim_{\mu \to 0^+} \theta(\mu) = \inf_{\mu > 0} \theta(\mu).
\]
Letting \(\theta(\mu) = f(x_\mu) + \mu B(x_\mu)\), where \(x_\mu \in X\) and \(g(x_\mu) < 0\), the limit of
any convergent subsequence of \(\{x_k\}\) is an optimal solution to the primal
problem, and furthermore, \(\mu B(x_\mu) \to 0\) as \(\mu \to 0^+\).

Proof

Let \(\bar{x}\) be an optimal solution to the primal problem satisfying the stated
property, and let \(\varepsilon > 0\). By the continuity of \(f\) and by the assumption of the
theorem, there is an \(\tilde{x} \in X\) with \(g(\tilde{x}) < 0\) such that \(f(\bar{x}) + \varepsilon > f(\tilde{x})\). Then, for \(\mu > 0\),
\[
f(\bar{x}) + \varepsilon + \mu B(\tilde{x}) > f(\tilde{x}) + \mu B(\tilde{x}) \geq \theta(\mu).
\]

\* Assumptions under which such a point \(x_\mu\) exists are given in Lemma 9.4.2.
Taking the limit as $\mu \to 0^+$, it follows that $f(\bar{x}) + \varepsilon \geq \lim_{\mu \to 0^+} \vartheta(\mu)$. Since this inequality holds true for each $\varepsilon > 0$, we get $f(\bar{x}) \geq \lim_{\mu \to 0^+} \vartheta(\mu)$. In view of Part 2 of Lemma 9.4.2, $f(\bar{x}) = \lim_{\mu \to 0^+} \vartheta(\mu)$.

For $\mu \to 0^+$ and since $B(x_\mu) \geq 0$ and $x_\mu$ is feasible to the original problem, it follows that

$$\vartheta(\mu) = f(x_\mu) + \mu B(x_\mu) \geq f(x_\mu) \geq f(\bar{x}).$$

Now taking the limit as $\mu \to 0^+$ and noting that $f(\bar{x}) = \lim_{\mu \to 0^+} \vartheta(\mu)$, it follows that both $f(x_\mu)$ and $f(x_\mu) + \mu B(x_\mu)$ approach $f(\bar{x})$. Therefore, $\mu B(x_\mu) \to 0$ as $\mu \to 0^+$. Furthermore, if $\{x_\mu\}$ has a convergent subsequence with limit $x'$, then $f(x') = f(\bar{x})$. Since $x_\mu$ is feasible to the original problem for each $\mu$, it follows that $x'$ is also feasible and hence optimal. This completes the proof.

Note that the points $\{x_\mu\}$ generated belong to the interior of the set $\{x : g(x) \leq 0\}$ for each $\mu$. It is for this reason that barrier function methods are sometimes also referred to as interior penalty function methods.

**KKT Lagrange Multipliers at Optimality**

Under certain regularity conditions, the barrier interior penalty method also produces a sequence of Lagrange multiplier estimates that converge to an optimal set of Lagrange multipliers. To see this, consider Problem P, to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$, and $x \in X = R^n$. (The case where $X$ might include additional inequality or equality constraints is easily treated in a similar fashion; see Exercise 9.19.) The barrier function problem is then given by

$$\min_{x} \left\{ f(x) + \mu \sum_{i=1}^{m} \phi(g_i(x)) : g(x) < 0 \right\}, \quad (9.32)$$

where $\phi$ satisfies (9.28). Let us assume that $f$, $g$, and $\phi$ are continuously differentiable, that the conditions of Lemma 9.4.2 and Theorem 9.4.3 hold true and that the optimum $\bar{x}$ to P obtained as an accumulation point of $\{x_\mu\}$ is a regular point. Without loss of generality, assume that $\{x_\mu\} \to \bar{x}$ itself. Then, if $I = \{i : g_i(\bar{x}) = 0\}$ is the index set of active constraints at $\bar{x}$, we know that there exists a unique set of Lagrange multipliers $\bar{\lambda}_i, i = 1, \ldots, m$, such that
\[
\n\n\n\n
Now since \( x_\mu \) solves the problem (9.32) with \( g(x_\mu) < 0 \), we have, for all \( \mu > 0 \),

\[
\n\n
As \( \mu \to 0^+ \) we have that \( \{x_\mu\} \to \bar{x} \), so \( (u_\mu)_i \to 0 \) for \( i \not\in I \). Moreover, since \( \bar{x} \) is regular and all the functions \( f, g, \) and \( \phi \) are continuously differentiable, we have, from (9.33) and (9.34), that \( (u_\mu)_i \to \bar{u}_i \) for \( i \in I \) as well. Hence, \( u_\mu \) provides an estimate for the set of Lagrange multipliers that approaches the optimal set of Lagrange multipliers \( \bar{u} \) as \( \mu \to 0^+ \). Therefore, for example, if \( \phi(y) = -1/y \), we have \( \phi'(y) = 1/y^2 \); hence,

\[
(u_\mu)_i = \frac{\mu}{g_i(x_\mu)^2} \bar{u}_i \quad \text{for all } i = 1, \ldots, m, \text{ as } \mu \to 0^+. \quad (9.35)
\]

Computational Difficulties Associated with Barrier Functions

The use of barrier functions for solving constrained nonlinear programming problems also faces several computational difficulties. First, the search must start with a point \( x \in X \) with \( g(x) < 0 \). For some problems, finding such a point may not be an easy task. In Exercise 9.24, a procedure is described for finding such a starting point. Also, because of the structure of the barrier function \( B \), and for small values of the parameter \( \mu \), most search techniques may face serious ill-conditioning and difficulties with round-off errors while solving the problem to minimize \( f(x) + \mu B(x) \) over \( x \in X \), especially as the boundary of the region \( \{x : g(x) \leq 0\} \) is approached. In fact, as the boundary is approached, and since search techniques often use discrete steps, a step leading outside the region \( \{x : g(x) \leq 0\} \) may indicate a decrease in the value of \( f(x) + \mu B(x) \), a false success. Thus, it becomes necessary to explicitly check the value of the constraint function \( g \) to guarantee that we do not leave the feasible region.

To see the potential ill-conditioning effect more formally, we can examine the eigenstructure of the Hessian of the objective function in (9.32) at the optimum \( x_\mu \) as \( \mu \to 0^+ \). Noting (9.34), and assuming that \( f, g, \) and \( \phi \) are twice continuously differentiable, we get that this Hessian is given by

\[
\begin{bmatrix}
V^2 f(x_\mu) + \sum_{i=1}^{m} (u_\mu)_i V^2 g_i(x_\mu) \\
\end{bmatrix} + \mu \sum_{i=1}^{m} \phi'[g_i(x_\mu)] V g_i(x_\mu) V g_i(x_\mu)' \quad (9.36)
\]

\[
\]
As \( \mu \to 0^+ \), we have \( \{x_\mu\} \to \bar{x} \) (possibly over a convergent subsequence); and assuming that \( \bar{x} \) is a regular point, we have \( u_\mu \to \bar{u} \), the optimal set of Lagrange multipliers. Hence, the term within \([\cdot]\) in (9.36) approaches \( \nabla^2 L(\bar{x}) \). The remaining term is potentially problematic. For example, if \( \phi(y) = -1/y \), then \( \phi''(y) = -2/y^3 \), so, from (9.35), this term becomes
\[
-2 \sum_{i=1}^{n} \frac{(u_\mu)_i \nabla g_i(x_\mu) \nabla g_i(x_\mu)^T}{g_i(x_\mu)}
\]
leading to an identical severe ill-conditioning effect as described for the exterior penalty functions. Hence, it is again imperative to use suitable second-order Newton, quasi-Newton, or conjugate gradient methods for solving the Problem (9.32).

**Summary of Barrier Function Methods**

We describe below a scheme using barrier functions for optimizing a nonlinear programming problem of the form to minimize \( f(x) \) subject to \( g(x) \leq 0 \) and \( x \in X \). The barrier function \( B \) used must satisfy (9.28).

The problem stated at Step 1 below incorporates the constraint \( g(x) < 0 \). If \( g(x_k) < 0 \), and since the barrier function approaches infinity as the boundary of the region \( G = \{x : g(x) < 0\} \) is reached, the constraint \( g(x) < 0 \) may be ignored, provided that an unconstrained optimization technique is used that will ensure that the resulting optimal point \( x_{k+1} \in G \). However, as most line search methods use discrete steps, if we are close to the boundary, a step could lead to a point outside the feasible region where the value of the barrier function \( B \) is a large negative number. Therefore, the problem could be treated as an unconstrained optimization problem only if an explicit check for feasibility is made.

**Initialization Step** Let \( \varepsilon > 0 \) be a termination scalar, and choose a point \( x_1 \in X \) with \( g(x_1) < 0 \). Let \( \mu_1 > 0, \beta \in (0, 1) \), let \( k = 1 \), and go to the Main Step.

**Main Step**

1. Starting with \( x_k \), solve the following problem:
   Minimize \( f(x) + \mu_k B(x) \)
   subject to \( g(x) < 0 \)
   \( x \in X \).

   Let \( x_{k+1} \) be an optimal solution, and go to Step 2.

2. If \( \mu_k B(x_{k+1}) < \varepsilon \), stop. Otherwise, let \( \mu_{k+1} = \beta \mu_k \), replace \( k \) by \( k + 1 \), and go to Step 1.
9.4.4 Example

Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad (x_1 - 2)^4 + (x_1 - 2x_2)^2 \\
\text{subject to} & \quad x_1^2 - x_2 \leq 0.
\end{align*}
\]

Here \( X = \mathbb{R}^2 \). We solve the problem by using the barrier function method with \( B(x) = -1/(x_1^2 - x_2) \). A summary of the computations is presented in Table 9.2, along with the Lagrange multiplier estimates as given by (9.35), and the progress of the algorithm is depicted in Figure 9.9. The procedure is started with \( \mu_1 = 10.0 \), and the unconstrained minimization of the function \( \theta(\mu_1) \) was started from a feasible point \((0.0, 1.0)\). The parameter \( \beta \) is taken as 0.10. After six iterations, the point \( x_7^* = (0.94389, 0.89635) \) and \( u_\mu = 3.385 \) is reached, where \( \mu_6 B(x_7) = 0.0184 \) and the algorithm is terminated. The reader can verify that this point is very close to the optimum. Noting that \( \mu_k \) is decreasing, the reader can observe from Table 9.2 that \( f(x_{\mu_k}) \) and \( \theta(\mu_k) \) are nondecreasing functions of \( \mu_k \). Similarly, \( B(x_{\mu_k}) \) is a nonincreasing function of \( \mu_k \). Furthermore, \( \mu_k B(x_{\mu_k}) \) converges to zero as asserted in Theorem 9.4.3.

9.5 Polynomial-Time Interior Point Algorithms for Linear Programming Based on a Barrier Function

Consider the following pair of primal (P) and dual (D) linear programming problems (see Section 2.7):

<table>
<thead>
<tr>
<th>Iter. ( k )</th>
<th>( \mu_k )</th>
<th>( x_{\mu_k} = x_{k+1} )</th>
<th>( f(x_{k+1}) )</th>
<th>( B(x_{k+1}) )</th>
<th>( \theta(\mu_k) )</th>
<th>( \mu_k B(x_{\mu_k}) )</th>
<th>( u_{\mu_k} )</th>
</tr>
</thead>
</table>
| 1            | 10.0        | \[
0.7079 \quad 1.5315
\]
|              |             | 8.3338          | 0.9705      | 18.0388     | 9.705        | 9.419051        |
| 2            | 1.0         | \[
0.8282 \quad 1.1098
\]
|              |             | 3.8214          | 2.3591      | 6.1805      | 2.3591       | 5.565503        |
| 3            | 0.1         | \[
0.8989 \quad 0.9638
\]
|              |             | 2.5282          | 6.4194      | 3.1701      | 0.6419       | 4.120815        |
| 4            | 0.01        | \[
0.9294 \quad 0.9162
\]
|              |             | 2.1291          | 19.0783     | 2.3199      | 0.1908       | 3.639818        |
| 5            | 0.001       | \[
0.9403 \quad 0.9011
\]
|              |             | 2.0039          | 59.0461     | 2.0629      | 0.0590       | 3.486457        |
| 6            | 0.0001      | \[
0.94389 \quad 0.89635
\]
|              |             | 1.9645          | 184.4451    | 1.9829      | 0.0184       | 3.385000        |
Figure 9.9 Barrier function method.

\[ \text{P: Minimize } c'x \quad \text{D: Maximize } b'v \]

subject to \[ Ax = b \quad \text{subject to } A'v + u = c \]
\[ x \geq 0, \quad u \geq 0, \quad v \text{ unrestricted,} \]

where \( A \) is an \( m \times n \) matrix and, without loss of generality, has rank \( m < n \), and where \( v \) and \( u \) are Lagrange multipliers associated with the equality and the inequality constraints of \( P \), respectively. Let us assume that \( P \) has an optimal solution \( x^* \), and let the corresponding optimal Lagrange multipliers be \( v^* \) and \( u^* \). Denoting by \( w \) the triplet \((x, u, v)\), we have that \( w^* = (x^*, u^*, v^*) \) satisfies the following KKT conditions for Problem \( P \):

\begin{align*}
Ax &= b, \quad x \geq 0 \quad (9.37a) \\
A'v + u &= c, \quad u \geq 0, \quad v \text{ unrestricted} \quad (9.37b) \\
u'x &= 0. \quad (9.37c)
\end{align*}

Now let us assume that there exists a \( w = (\bar{x}, \bar{u}, \bar{v}) \) satisfying (9.37a) and (9.37b) with \( \bar{x} > 0 \) and \( \bar{u} > 0 \). Consider the following barrier function problem \( \text{BP} \) based on Frisch's logarithmic barrier function (9.29b), where the equality constraints are used to define the set \( X \):
BP: Minimize \[
\begin{align*}
\mathbf{c}'\mathbf{x} - \mu \sum_{j=1}^{n} \ln(x_j) : \mathbf{A}\mathbf{x} = \mathbf{b}, (\mathbf{x} > 0).
\end{align*}
\] (9.38)

The KKT conditions for BP require that we find \(\mathbf{x}\) and \(\mathbf{v}\) such that \(\mathbf{A}\mathbf{x} = \mathbf{b}\) \((\mathbf{x} > 0)\) and \(\mathbf{A}'\mathbf{v} = \mathbf{c} - \mu[1/x_1, ..., 1/x_n]'\). Following (9.34), we can denote \(\mathbf{u} = \mu[1/x_1, ..., 1/x_n]'\) as our Lagrange multiplier estimate for \(P\) given any \(\mu > 0\). Defining the diagonal matrices \(\mathbf{X} = \text{diag}\{x_1, ..., x_n\}\) and \(\mathbf{U} = \text{diag}\{u_1, ..., u_n\}\), and denoting \(\mathbf{e} = (1, ..., 1)'\) as a conformable vector of ones we can rewrite the KKT conditions for BP as follows:

\[
\begin{align*}
\mathbf{A}\mathbf{x} &= \mathbf{b} \quad \text{(9.39a)} \\
\mathbf{A}'\mathbf{v} + \mathbf{u} &= \mathbf{c} \quad \text{(9.39b)} \\
\mathbf{u} &= \mu\mathbf{X}^{-1}\mathbf{e} \quad \text{or} \quad \mathbf{X}\mathbf{U}\mathbf{e} = \mu\mathbf{e}. \quad \text{(9.39c)}
\end{align*}
\]

Note that the innocuous alternative equivalent form \(\mathbf{X}\mathbf{U}\mathbf{e} = \mu\mathbf{e}\) used in Equation (9.39c) plays a key role in the subsequent application of Newton's method, yielding an algorithmic behavior that is not reproducible by applying an identical strategy to the original form of this equation, \(\mathbf{u} = \mu\mathbf{X}^{-1}\mathbf{e}\), given in (9.39c). Hence, the system (9.39) with \(\mathbf{X}\mathbf{U}\mathbf{e} = \mu\mathbf{e}\) used in (9.39c) is sometimes referred to as the perturbed KKT system for Problem BP.

Now, given any \(\mu > 0\), by Part 1 of Lemma 9.4.2 and by the strict convexity of the objective function of Problem BP over the feasible region, there exists an unique \(\mathbf{x}_\mu > 0\) that solves BP. Correspondingly, from (9.39), since \(\mathbf{A}'\) has full column rank, we obtain unique accompanying values of \(\mathbf{u}_\mu\) and \(\mathbf{v}_\mu\). Following Theorem 9.4.3, we can show that the triplet \(\mathbf{w}_\mu \equiv (\mathbf{x}_\mu, \mathbf{u}_\mu, \mathbf{v}_\mu)\) approaches an optimal primal-dual solution to \(P\) as \(\mu \to 0^+\). The trajectory \(\mathbf{w}_\mu\) for \(\mu > 0\) is known as the central path because of the interiority forced by the barrier function. Note that from (9.39 a, b), the standard linear programming duality gap \(\mathbf{c}'\mathbf{x} - \mathbf{b}'\mathbf{v}\) equals \(\mathbf{u}'\mathbf{x}\), the total violation in the complementary slackness condition. Moreover, from (9.39c), we get \(\mathbf{u}'\mathbf{x} = \mu\mathbf{x}'\mathbf{X}^{-1}\mathbf{e} = n\mu\). Hence, we have, based on (9.39), that

\[
\mathbf{c}'\mathbf{x} - \mathbf{b}'\mathbf{v} = \mathbf{u}'\mathbf{x} = n\mu,
\] (9.40)

which approaches zero as \(\mu \to 0^+\).

Instead of actually finding \(\mathbf{w}_\mu\) for each \(\mu > 0\) in a sequence approaching zero, we start with a \(\bar{\mu} > 0\) and a \(\bar{\mathbf{w}}\) sufficiently close to \(\mathbf{w}_\bar{\mu}\) and then revise \(\bar{\mu}\) to \(\bar{\mu} = \beta\bar{\mu}\) for some \(0 < \beta < 1\). Correspondingly, we shall then use a single
Newton step to obtain a revised solution \( \hat{w} \) that is also sufficiently close to \( w_{\mu} \).
Motivated by (9.39) and (9.40), by defining \( w = (x, u, v) \) to be "sufficiently close" to \( w_{\mu} \) whenever

\[
Ax = b, \quad A'v + u = c, \quad \|Xu - \mu e\| \leq \theta \mu
\]

with \( u'x = n\mu \), where \( 0 \leq \theta < 0.5 \)

we shall show that such a sequence of iterates \( w \) will then converge to an optimal primal-dual solution.

Toward this end, suppose that we are given a \( \bar{\mu} > 0 \) and a \( \bar{w} = (\bar{x}, \bar{u}, \bar{v}) \) with \( \bar{x} > 0 \) and \( \bar{u} > 0 \), such that (9.41) holds true. (Later, we show how one can obtain such a solution to initialize the algorithm.) Let us now reduce \( \mu \) to \( \mu = \beta \bar{\mu} \), where \( 0 < \beta < 1 \), and examine the perturbed KKT system (9.39) written for \( \mu = \bar{\mu} \). Denote this system as \( H(w) = 0 \). The first-order approximation of this equation system at \( w = \bar{w} \) is given by \( H(\bar{w}) + J(\bar{w})(w - \bar{w}) = 0 \), where \( J(\bar{w}) \) is the Jacobian of \( H(w) \) at \( w = \bar{w} \). Denoting \( d_w = (w - \bar{w}) \), a Newton step at \( \bar{w} \) will take us to the point \( \hat{w} = \bar{w} + d_w \) where \( J(\bar{w})d_w = -H(\bar{w}) \). Writing \( d_w = (d_x, d_u', d_v') \), we have from (9.39) that the equation \( J(\bar{w})d_w = -H(\bar{w}) \) is given as follows:

\[
\begin{align*}
Ad_x &= 0 \quad (9.42a) \\
A'd_v + d_u &= 0 \quad (9.42b) \\
\bar{U}d_x + \bar{X}d_u &= \mu e - \bar{X}ue. \quad (9.42c)
\end{align*}
\]

The linear system (9.42) can be solved using some stable, factored-form implementation (see Appendix A.2). In explicit form we get \( d_u = -A'd_v \) from (9.42b) and hence \( d_x = \bar{U}^{-1}[\mu e - \bar{X}ue] + \bar{U}^{-1}\bar{X}A'd_v \). Substituting this in (9.42a) gives

\[
\begin{align*}
d_v &= -[A\bar{U}^{-1}\bar{X}A']^{-1}A\bar{U}^{-1}[\mu e - \bar{X}ue] \quad (9.43a) \\
d_u &= -A'd_v \quad (9.43b) \\
d_x &= \bar{U}^{-1}[\mu e - \bar{X}ue - \bar{X}d_u], \quad (9.43c)
\end{align*}
\]

where the inverses exist, since \( \bar{x} > 0, \bar{u} > 0 \), and \( \text{rank}(A) = m \). The generation of \((\hat{\mu}, \hat{w})\) from \((\bar{\mu}, \bar{w})\) in the above fashion describes one step of the algorithm.

This procedure can now be repeated until the duality gap \( u'x = n\mu \) [see (9.37),
(9.40), and (9.41)] is small enough. The algorithmic steps are summarized below. This algorithm is called a \textit{path-following procedure} because of its attempt to (approximately) follow the central path.

\textbf{Summary of the Primal-Dual Path-Following Algorithm}

\textbf{Initialization} Select a starting solution \( \bar{\mathbf{w}} = (\bar{\mathbf{x}}, \bar{\mathbf{u}}, \bar{\mathbf{v}}) \) with \( \bar{\mathbf{x}} > 0, \bar{\mathbf{u}} > 0 \), and a penalty parameter \( \mu = \bar{\mu} \) such that \( \bar{\mathbf{w}} \) satisfies (9.41) with \( \mu = \bar{\mu} \). (Later we show how this can be accomplished for a general linear program.) Furthermore, let \( \theta, \delta, \) and \( \beta \) satisfy (9.44) of Theorem 9.5.2 stated below \([e.g., \text{let } \theta = \delta = 0.35 \text{ and } \beta = 1 - (\delta/\sqrt{n})]\). Put \( k = 0, \) let \( (\mu_0, w_0) = (\bar{\mu}, \bar{\mathbf{w}}) \), and proceed to the Main Step.

\textbf{Main Step} Let \( (\bar{\mu}, \bar{\mathbf{w}}) = (\mu_k, w_k) \). If \( \mathbf{c}^T \bar{\mathbf{x}} - \mathbf{b}^T \bar{\mathbf{v}} = n\bar{\mu} < \varepsilon \) for some tolerance \( \varepsilon > 0 \), then terminate with \( \bar{\mathbf{w}} \) as an \( (\varepsilon-) \) optimal primal-dual solution.

Otherwise, let \( \bar{\mu} = \beta \bar{\mu} \), compute \( \mathbf{d}_w^T = (\mathbf{d}_x^T, \mathbf{d}_u^T, \mathbf{d}_v^T) \) by (9.43) \([\text{or through } (9.42)]\), and set \( \bar{\mathbf{w}} = \bar{\mathbf{w}} + \mathbf{d}_w \). Put \( (\mu_{k+1}, w_{k+1}) = (\bar{\mu}, \bar{\mathbf{w}}) \), increment \( k \) by 1, and repeat the Main Step.

\textbf{9.5.1 Example}

Consider the linear program to minimize \( \{3x_1 - x_2 : x_1 + 2x_2 = 2, x_1 \geq 0, x_2 \geq 0\} \). The optimal solution is easily seen to be \( \mathbf{x}^* = (0, 1)^T \), with \( \mathbf{v}^* = -0.5 \) and \( \mathbf{u}^* = (3.5, 0)^T \). Suppose that we initialize the algorithm with \( \bar{\mathbf{x}} = (2/9, 8/9)^T \), \( \bar{\mathbf{u}} = (4, 1)^T \), \( \bar{\mathbf{v}} = -1 \), and \( \bar{\mu} = 8/9 \). This solution can be verified to satisfy (9.39) and, therefore, \( \bar{\mathbf{w}} = w_{\bar{\mu}} \) lies on the central path. Hence, in particular, \( (\mu_0, w_0) = (\bar{\mu}, \bar{\mathbf{w}}) \) satisfies (9.41). The present duality gap from (9.40) is given by \( \bar{\mathbf{u}}^T \bar{\mathbf{x}} = 2\bar{\mu} = 1.777777777 \). Let \( \theta = \delta = 0.35 \) and \( \beta = 1 - (0.35/\sqrt{2}) = 0.7525127 \).

Now let us compute \( (\mu_1, w_1) \equiv (\bar{\mu}, \bar{\mathbf{w}}) \) according to \( \bar{\mu} = \beta \bar{\mu} \) and \( \bar{\mathbf{w}} = \bar{\mathbf{w}} + \mathbf{d}_w \), where \( \mathbf{d}_w^T = (\mathbf{d}_x^T, \mathbf{d}_u^T, \mathbf{d}_v^T) \) is the solution to (9.42). Hence, we obtain \( \mu_1 = \bar{\mu} = 0.6689001 \), and \( \mathbf{d}_w \) solves the system

\begin{align*}
    d_{x_1} + 2d_{x_2} &= 0 \\
    d_{u_1} + d_{u_2} &= 0 \\
    2d_{v_1} + d_{u_2} &= 0 \\
    4d_{x_1} + 2d_{u_1} &= \bar{\mu} - \bar{x}_1\bar{u}_1 = -0.2199887 \\
    d_{x_2} + d_{u_2} &= \bar{\mu} - \bar{x}_2\bar{u}_2 = -0.2199887.
\end{align*}
Solving, we obtain \( d_v = 0.1370698 \), \( d_u = (-0.1370698, -0.2741396)' \), and \( d_x = (-0.0473822, 0.0236909)' \). This yields \( w_1 = \hat{w} = (\hat{x}, \hat{u}, \hat{v}) = \bar{w} + d_x \), where \( \hat{x} = (0.17484, 0.9125797)' \), \( \hat{u} = (3.8629302, 0.7258604)' \), and \( \hat{v} = -0.8629302 \).

Note that the duality gap has reduced to \( \hat{u}'\hat{x} = 1.3378 = 2\hat{\mu} \). Also, observe that \( \hat{X}\hat{U}e = (\hat{x}_1\hat{u}_1, \hat{x}_2\hat{u}_2)' = (0.6753947, 0.6624054)' \neq \hat{\mu}e \), and hence, (9.39c) no longer holds true and we are not located on the central path. However, \( \hat{w} \) is close enough to \( w_{\bar{\mu}} \) in the sense of (9.41), since \( \|\hat{X}\hat{U}e - \hat{\mu}e\| = 0.009176 \leq \theta\hat{\mu} = 0.234115 \).

We ask the reader in Exercise 9.32 to continue the iterations until (near) optimality is attained. Below, we establish the main result.

9.5.2 Theorem

Let \( \bar{w} = (\bar{x}, \bar{u}, \bar{v}) \) be such that \( \bar{x} > 0, \bar{u} > 0 \), and (9.41) is satisfied with \( \mu = \bar{\mu} \). Consider \( \hat{\mu} = \beta\bar{\mu} \), where \( 0 < \beta < 1 \) satisfies

\[
\beta = 1 - \frac{\delta}{\sqrt{n}}
\]

and where

\[
0 < \delta < \sqrt{n}, \quad \frac{\theta^2 + \delta^2}{2(1-\theta)} \leq \theta \left(1 - \frac{\delta}{\sqrt{n}}\right), \quad 0 \leq \theta \leq \frac{1}{2}.
\]

(For example, we can take \( \theta = \delta = 0.35 \).) Then the solution \( \hat{w} = \bar{w} + d_w \) produced by taking a unit step length along the Newton direction \( d_w \) given by (9.43) [or (9.42)] has \( \hat{x} > 0, \hat{u} > 0 \), and also satisfies (9.41) with \( \mu = \hat{\mu} \). Hence, starting with \((\mu_0, w_0)\) satisfying (9.41), the algorithm generates a sequence \( \{(\mu_k, w_k)\} \) that satisfies (9.41) at each iteration, and such that any accumulation point of \( \{w_k\} \) solves the original linear program.

Proof

First, note from (9.42a) and (9.42b) that

\[
\mathbf{A}\hat{x} = \mathbf{A}\bar{x} = \mathbf{b} \quad \text{and} \quad \mathbf{A}'\hat{v} + \hat{u} = \mathbf{A}'\bar{v} + \bar{u} = \mathbf{c}.
\]

Now let us show that \( \|\hat{X}\hat{U}e - \hat{\mu}e\| \leq \theta\hat{\mu} \). Denoting \( \mathbf{D}_x = \text{diag}\{d_{x_1}, \ldots, d_{x_n}\} \), \( \mathbf{D}_u = \text{diag}\{d_{u_1}, \ldots, d_{u_n}\} \), and \( \mathbf{D} = (\bar{X}^{-1}\bar{U})^{1/2} \), we have from (9.42c) that

\[
\hat{X}\hat{U}e - \hat{\mu}e = (\bar{X} + \mathbf{D}_x)(\bar{U} + \mathbf{D}_u)e - \hat{\mu}e = \mathbf{D}_x\mathbf{D}_ue.
\]

Moreover, multiplying (9.42c) throughout by \( (\bar{X}\bar{U})^{-1/2} \), we get
Penalty and Barrier Functions

\[ DD_x \epsilon + D^{-1} D_u \epsilon = (\bar{X} \bar{U})^{-1/2} [\hat{\mu} \epsilon - \bar{X} \bar{U} \epsilon]. \]  

(9.47)

Hence,

\[ \| \hat{X} \epsilon - \hat{\mu} \epsilon \| = \| D_x D_u \epsilon \| = \| (DD_x)(D^{-1} D_u) \epsilon \| = \sqrt{\sum \left( \pi_j \gamma_j \right)^2}, \]

where \( DD_x = \text{diag}\{\pi_1, \ldots, \pi_n\} \), say, and \( D^{-1} D_u = \text{diag}\{\gamma_1, \ldots, \gamma_n\} \), say. Using (9.47) and denoting \( \bar{x}_u_{\text{min}} = \min\{x_j \bar{u}_j, j = 1, \ldots, n\} \), we get

\[ \| \hat{X} \epsilon - \hat{\mu} \epsilon \| \leq \sum \pi_j \gamma_j \leq \frac{1}{2} \sum (\pi_j + \gamma_j)^2 = \frac{1}{2} \| DD_x \epsilon + D^{-1} D_u \epsilon \|^2, \]

(9.48)

\[ \| \hat{X} \epsilon - \hat{\mu} \epsilon \| ^2 = \| (X \epsilon - \bar{X} \epsilon) \|^2 = \sum \frac{(\hat{\mu} - x_j \bar{u}_j)^2}{x_j \bar{u}_j} \leq \frac{\sum (\hat{\mu} - x_j \bar{u}_j)^2}{2 \bar{x}_u_{\text{min}}} \]

(9.49)

Furthermore, from (9.41), since \( \| \hat{X} \epsilon - \hat{\mu} \epsilon \| \leq \theta \bar{\mu} \) implies that \( |x_j \bar{u}_j - \bar{\mu}| \leq \theta \bar{\mu} \), we get that \( \bar{\mu} - x_j \bar{u}_j \leq \theta \bar{\mu} \) for all \( j = 1, \ldots, n \), or that

\[ \bar{x}_j \bar{u}_j \geq \bar{\mu}(1 - \theta) \quad \text{for all} \quad j = 1, \ldots, n, \quad \text{so} \quad \bar{x}_u_{\text{min}} \geq \bar{x} \bar{\mu}(1 - \theta). \]

(9.50)

Using (9.49) and (9.50) in (9.48), and noting (9.44), we derive

\[ \| \hat{X} \epsilon - \hat{\mu} \epsilon \| \leq \frac{\theta \bar{\mu} \left[ \theta^2 + n(1 - \beta)^2 \right]}{2(1 - \theta)} = \frac{\theta \bar{\mu} \left[ \bar{\theta}^2 + \gamma^2 \right]}{2(1 - \theta)} \leq \theta \bar{\mu} = \theta \hat{\mu}. \]

(9.51)

Hence, we have \( \| \hat{X} \epsilon - \hat{\mu} \epsilon \| \leq \theta \hat{\mu} \). Let us now show that \( \hat{u} \hat{x} = n \hat{\mu} \). Using \( \hat{w} = \bar{w} + d_w \), we obtain

\[ \hat{u} \hat{x} = (\bar{u} + d_u)'(\bar{x} + d_x) = \bar{u}' \bar{x} + \bar{u}' d_x + d_u' \bar{x} + d_u' d_x = e' [\bar{X} \epsilon + \bar{U} d_x + \bar{X} d_u] + d_u' d_x. \]

\[ = e' [\bar{X} \epsilon + \bar{U} d_x + \bar{X} d_u] + d_u' d_x. \]
From (9.42c), the term in \([\cdot]\) equals \(\hat{\mu}e\). Furthermore, from (9.42a, b), we observe that \(d_u' d_x = -d_x' A d_x = 0\). Hence, this gives \(\hat{u}' \hat{x} = e'(\hat{\mu}e) = n\hat{\mu}\), and this, along with (9.45) and (9.51), shows that \(\hat{w}\) satisfies (9.41) with \(\mu = \hat{\mu}\).

To complete the proof of the first assertion of the theorem, we now need to show that \(\hat{x} > 0\) and \(\hat{u} > 0\). Toward this end, following (9.50) stated for \(\hat{w}\) and \(\hat{\mu}\), we have

\[
\hat{x}_j \hat{u}_j \geq \hat{\mu}(1 - \theta) > 0 \quad \text{for all } j = 1, \ldots, n. \tag{9.52}
\]

Hence, for each \(j = 1, \ldots, n\), either \(\hat{x}_j > 0\) and \(\hat{u}_j > 0\), or else \(\hat{x}_j < 0\) and \(\hat{u}_j < 0\). Assuming the latter for some \(j\), on the contrary, since \(\hat{x}_j = \bar{x}_j + (d_x)_j\) and \(\hat{u}_j = \bar{u}_j + (d_u)_j\), where \(\bar{x}_j > 0\) and \(\bar{u}_j > 0\), we have that \((d_x)_j < \hat{x}_j < 0\) and \((d_u)_j < \hat{u}_j < 0\); so, from (9.52), we obtain

\[
(d_x)_j (d_u)_j > \hat{x}_j \hat{u}_j \geq \hat{\mu}(1 - \theta). \tag{9.53}
\]

But from (9.46) and (9.51), we have

\[
(d_x)_j (d_u)_j \leq \|D_x D_u e\| = \| \hat{X} \hat{U} e - \hat{\mu}e \| \leq \hat{\mu} \theta. \tag{9.54}
\]

Equations (9.53) and (9.54) imply that \(\hat{\mu}(1 - \theta) < \hat{\mu} \theta\), or that \(\theta > 0.5\), which contradicts (9.44). Hence, \(\hat{x} > 0\) and \(\hat{u} > 0\).

Finally, observe from (9.41) and the foregoing argument that the algorithm generates a sequence \(w_k = (x_k, u_k, v_k)\) and a sequence \(\mu_k\) such that

\[
A x_k = b, \ x_k > 0, \ A' v_k + u_k = c, \ u_k > 0, \ u_k' x_k = n \mu_k = n \mu_0 (\beta)^k. \tag{9.55}
\]

Since \(\beta^k \to 0\) as \(k \to \infty\), any accumulation point \(w^* = (x^*, u^*, v^*)\) of the sequence \(\{w_k\}\) satisfies the necessary and sufficient optimality conditions (9.37) for Problem P and, hence, yields a primal-dual optimal solution to P. This completes the proof.

**Convergence Rate and Complexity Analysis**

Observe from (9.40) and (9.55) that the duality gap \(c' x_k - b' v_k\) for the pair of primal-dual feasible solutions \(x_k > 0\) and \(v_k\) (with slacks \(u_k > 0\)) generated by the algorithm equals \(u_k' x_k = n \mu_k = n \mu_0 (\beta)^k\) and approaches zero at a geometric (linear) rate of convergence. Moreover, the convergence rate ratio \(\beta\) is given by \(\beta = 1 - (\delta/\sqrt{n})\), which for a fixed value of \(\delta\), approaches unity as \(n \to \infty\), implying an increasingly slower convergence rate behavior as \(n\) increases. Hence, from a practical standpoint, implementations of this algorithm tend to
shrink \( \mu \) faster to zero based on the duality gap [e.g., by taking \( \mu_{k+1} = (c'x_k - b'v_k) / \zeta(n) \), where \( \zeta(n) = n^2 \) for \( n \leq 5000 \) and \( \zeta(n) = n\sqrt{n} \) for \( n \geq 5000 \)] and also to conduct line searches along the Newton direction \( d_w \) while maintaining \( x_k > 0 \) and \( u_k > 0 \) rather than simply taking a unit step size.

The unmodified form of the algorithm possesses the desirable property of having a polynomial-time complexity. This means that if the data for \( P \) is all integer, and if \( L \) denotes the number of binary bits required to represent the data, then while using a number of elementary operations (additions, multiplications, comparisons, etc.) bounded above by a polynomial in the size of the problem as defined by the parameters \( m \), \( n \), and \( L \), the algorithm will finitely determine an exact optimum to Problem \( P \). It can be shown (see the Notes and References section) that we can begin with \( \|x_0\| < 2^L \) and \( \|u_0\| < 2^L \) and that once the duality gap satisfies \( u_k x_k \leq 2^{-2L} \) for some \( k \), the solution, \( w_k \), can be purified (or rounded) in polynomial time to an exact optimum by a process that finds a vertex solution having at least as good an objective value as that obtained currently. Note that from (9.55) and (9.41) we have

\[
\begin{align*}
\mu_0(\beta)^k &= u_0 x_0(\beta)^k < 2^{2L}(\beta)^k \leq 2^{-2L}, \\
\text{when } \beta^k &\leq 2^{-4L} \quad \text{or} \quad k \geq \frac{[4 \ln(2)]L}{-\ln(\beta)}. \quad (9.56)
\end{align*}
\]

But by the concavity of \( \ln(\cdot) \) over the positive real line, we have from (9.44) that \( \ln(\beta) = \ln[1 - (\delta/\sqrt{n})] \leq -\delta/\sqrt{n} \); so \( \ln(\beta) \geq (\delta/\sqrt{n}) \). Consequently, when \( k \geq [4 \ln(2)L]/(\delta/\sqrt{n}) \), we have (9.56) holding true, and we can then purify the available solution finitely in polynomial time to an exact optimum. The number of iterations before this happens is therefore bounded above by a constant times \( \sqrt{n} L \); this is denoted as being of order of complexity \( O(\sqrt{n} L) \). Because each iteration itself requires a number of operations bounded above by a polynomial in \( n \) [refer to the Notes and References section on how this can be achieved in \( O(n^{2.5}) \) steps per iteration by a process of updating the solution to the system (9.42) from one iteration to the next], the overall algorithm is of polynomial-time complexity \( O(n^3 L) \). In the Notes and References section we refer the reader to a discussion on modifying the algorithm by adaptively varying the parameters from one iteration to the next so that superlinear convergence is realized, without impairing the polynomial-time behavior of the algorithm.

**Getting Started**

To complete our discussion of the primal-dual path following algorithm, let us show how we can initialize this procedure. Note that the given pair of primal and dual problems \( P \) and \( D \) need not possess a primal-dual feasible solution \( w = (x, u, v) \), with \( x > 0 \) and \( u > 0 \), and such that (9.41) holds true. Hence, we
Chapter 9

employ artificial variables as in Section 2.7 to accomplish this requirement. Toward this end, let \( \lambda \) and \( \gamma \) be sufficiently large scalars. (Theoretically, we can take \( \lambda = 2^{2L} \) and \( \gamma = 2^{4L} \), although practically, these values might be far too large.) Define

\[
M_1 = \lambda \gamma \quad \text{and} \quad M_2 = \lambda \gamma (n+1) - \lambda (e^t c).
\]

Let \( x_a \) be a single artificial variable, define an auxiliary variable \( x_{n+1} \), and consider the following (big-M) artificial primal problem \( P' \) along with its dual \( D' \) (see Section 2.7):

\[
\begin{align*}
P': & \quad \text{Minimize} & c'x + M_1 x_a \\
& \quad \text{subject to} & Ax + (b - \lambda Ae)x_a = b \\
& & (c - \gamma e)'x - \gamma x_{n+1} = -M_2 \\
& & (x, x_a, x_{n+1}) \geq 0. \\
D': & \quad \text{Maximize} & b'v - M_2 v_a \\
& \quad \text{subject to} & A'v + (c - \gamma e)v_a + u = c \\
& & (b - \lambda Ae)'v + u_a = M_1 \\
& & -\gamma v_a + u_{n+1} = 0 \\
& & (u, u_a, u_{n+1}) \geq 0 \\
& & (v, v_a) \text{ unrestricted.}
\end{align*}
\]

It is easily verified that the primal-dual pair of solutions

\[
(x', x_a, x_{n+1}) = (\lambda e^t, 1, \lambda) > 0
\]

and

\[
(9.58)
\]

are feasible to \( P' \) and \( D' \), respectively, and, moreover, that

\[
\begin{align*}
x_j u_j & = \lambda \gamma \quad \text{for} \quad j = 1, \ldots, n, \\
x_a u_a & = \lambda \gamma, \\
x_{n+1} u_{n+1} & = \lambda \gamma.
\end{align*}
\]

Consequently, with \( \mu \) initialized as \( \lambda \gamma \), the solution (9.58) lies on the central path, so (9.41) holds true. Hence, this solution can be used to initialize the algorithm to solve the pair of primal-dual problems \( P' \) and \( D' \). As with the artificial variable method discussed in Section 2.7, it can be shown that if \( x_a = v_a = 0 \) at optimality, then the corresponding optimal solutions \( x \) and \( v \) solve \( P \) and \( D \), respectively. Otherwise, at termination, if \( x_a > 0 \) and \( v_a = 0 \), then \( P \) is infeasible; if \( x_a = 0 \) and \( v_a > 0 \), then \( P \) is unbounded; and if \( x_a > 0 \) and \( v_a > 0 \), then \( P \) is either infeasible or unbounded. The last case can be resolved by
replacing P with its Phase I problem of Section 2.7. Also, since the size of P' is polynomially related to size of P, the polynomial-time complexity property of the algorithm is preserved.

**Predictor-Corrector Algorithms**

In concluding this section, we comment on a highly computationally effective interior point variant known as the predictor-corrector approach. The basic idea behind this technique originates from the successive approximation methods that are implemented for the numerical solution of differential equations. Essentially, this method adopts steps along two successive directions at each iteration. The first is a predictor step that adopts a direction based on the system (9.42) but under the ideal scenario of $\mu = 0$ used in this system. Let $d_w^*$ be this direction obtained. A tentative revised iterate $w'$ is computed as $w' = w + \lambda d_w^*$, where the step length $\lambda$ is taken close to the maximum step length $\lambda_{\text{max}}$ that would maintain nonnegativity of the $x$ and $u$ variables. (Typically, we take $\lambda = \alpha \lambda_{\text{max}}$ when $\alpha = 0.95-0.99$.) Then a revised value $\mu$ of the parameter $\mu$ is determined via a scheme of the type $\mu = \mu/n$ if $n \leq 5000$ and $\mu = \mu/\sqrt{n}$ if $n > 5000$, or based on some suitable function of the optimality gap or complementary slackness violation at $\overline{w}$ and at $w'$ [see Equation (9.40) and Exercise 9.31]. Using this $\mu$ value in the perturbed KKT system (9.39) and writing

$$x = \overline{x} + d_x, \quad u = \overline{u} + d_u, \quad \text{and} \quad v = \overline{v} + d_v$$

(9.59)

yields upon using that $\overline{w} = (\overline{x}, \overline{u}, \overline{v})$ satisfies (9.41) that

$$A_d x = 0$$

(9.60a)

$$A' d_v + d_u = 0$$

(9.60b)

$$\overline{U} d_x + \overline{X} d_u = \hat{\mu} e - \overline{X} e - D_x D_u e,$$

(9.60c)

where $D_x \equiv \text{diag}\{d_{x_1}, \ldots, d_{x_n}\}$ and $D_u \equiv \text{diag}\{d_{u_1}, \ldots, d_{u_n}\}$. Observe that if we drop the quadratic term $D_x D_u e = (d_{x_j} d_{u_j}, j = 1, \ldots, n)'$ in (9.60c), we obtain precisely the linearized system (9.42).

Now, in this system (9.60), we replace the quadratic term $D_x D_u e$ within (9.60c) by the estimate $(d'_{x_j} d'_{u_j}, j = 1, \ldots, n)'$ as determined by $d_w^*$ in lieu of simply neglecting this nonlinear term, and then solve this system to obtain the direction $d_w = (d_x, d_u, d_v)$. (Note that because of the similarity of the systems that determine $d_w^*$ and $d_w$, factorizations developed to compute the former direction can be reutilized to obtain the latter direction.) The revised iterate $\hat{w}$ is then computed as $\hat{w} = w + d_w$. This correction toward the central path is known as a corrector step. Observe that although the system (9.60) could be solved in
this manner repeatedly by using the most recent direction components to estimate the quadratic term on the right-hand side in (9.60c), this has not been observed to be computationally advisable. Hence, a single corrector step is adopted in practice. We refer the reader to the Notes and References section for further reading on this subject.

Exercises

[9.1] Given the set of inequality constraints \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \), any of the following auxiliary functions may be employed:

\[
\begin{align*}
  f(x) + \mu \sum_{i=1}^{m} \max\{0, g_i(x)\}, \\
  f(x) + \mu \sum_{i=1}^{m} [\max\{0, g_i(x)\}]^2, \\
  f(x) + \mu \max\{0, g_1(x), \ldots, g_m(x)\}, \\
  f(x) + \mu [\max\{0, g_1(x), \ldots, g_m(x)\}]^2.
\end{align*}
\]

Compare among these forms. What are the advantages and disadvantages of each?

[9.2] Consider the following problem:

Minimize \( 2e^x + 3x_1^2 + 2x_1x_2 + 4x_2^2 \)
subject to \( 3x_1 + 2x_2 - 6 = 0 \).

Formulate a suitable exterior penalty function with \( \mu = 10 \). Starting with the point \((1, 1)\), perform two iterations of some conjugate gradient method.

[9.3] This exercise describes several strategies for modifying the penalty parameter. Consider the following problem:

Minimize \( 2(x_1 - 3)^2 + (x_2 - 5)^2 \)
subject to \( 2x_1^2 - x_2 \leq 0 \).

Using the auxiliary function \( 2(x_1 - 3)^2 + (x_2 - 5)^2 + \mu \max\{2x_1^2 - x_2, 0\} \) and adopting the cyclic coordinate method, solve the above problem starting from the point \( x_1 = (0, -3)' \) under the following strategies for modifying \( \mu \):

a. Starting from \( x_1 \), solve the penalty problem for \( \mu_1 = 0.1 \) resulting in \( x_2 \). Then, starting from \( x_2 \), solve the problem with \( \mu_2 = 100 \).

b. Starting from the unconstrained optimal point \((3, 5)\), solve the penalty problem for \( \mu_2 = 100 \). (This is akin to Part a for \( \mu_1 = 0 \).)
c. Starting from $x_1$, apply the algorithm described in Section 9.2 by using the successively increasing values of $\mu = 0.1, 1.0, 10.0,$ and 100.0.

d. Starting from $x_1$, solve the penalty problem for $\mu_1 = 100.0$.

Which of the above strategies would you recommend, and why? Also, in each case above, derive an estimate for the Lagrange multiplier associated with the single constraint.

[9.4] Consider the following problem:

$$\text{Minimize } x_1^2 + 2x_2^2$$
$$\text{subject to } 2x_1 + 3x_2 - 6 \leq 0$$
$$-x_2 + 1 \leq 0.$$  

da. Find the optimal solution to this problem.
b. Formulate a suitable function, with an initial penalty parameter $\mu = 1$.
c. Starting from the point $(2, 4)$, solve the resulting problem by a suitable unconstrained minimization technique.
d. Replace the penalty parameter $\mu$ by 10. Starting from the point you obtained in Part c, solve the resulting problem.

[9.5] Consider the following problem:

$$\text{Minimize } 2x_1^2 - 3x_1x_2 + x_2^2$$
$$\text{subject to } x_1^2 - x_2 + 3 \leq 0$$
$$3x_1 + 2x_2 - 6 \leq 0$$
$$x_1, x_2 \geq 0.$$  

Solve the above problem by an exterior penalty function method starting from $(0, 0)$ for each specification of $X$:

a. $X = \mathbb{R}^n$.
b. $X = \{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0\}$.
c. $X = \{(x_1, x_2) : 3x_1 + 2x_2 - 6 \leq 0, x_1 \geq 0, x_2 \geq 0\}$.

(Effective methods for handling linear constraints are discussed in Chapter 10.) Compare among the above three alternative approaches. Which would you recommend?

[9.6] Consider the problem to minimize $x^3$ subject to $x = 1$. Obviously, the optimal solution is $\bar{x} = 1$. Now consider the problem to minimize $x^3 + \mu(x-1)^2$.

a. For $\mu = 1.0, 10.0, 100.0,$ and 1000.0, plot $x^3 + \mu(x-1)^2$ as a function of $x$, and for each case, find the point where the derivative
of the function vanishes. Also, verify that the optimal solution is unbounded.

b. Show that the optimal solution to the penalty problem is unbounded for any given \(\mu\), so that the conclusion of Theorem 9.2.2 does not hold true. Discuss.

c. For \(\mu = 1.0, 10.0, 100.0,\) and \(1000.0\), find the optimal solution to the penalty problem with the added constraint \(|x| \leq 2\).

[9.7] Consider the following problem:

\[
\text{Minimize } x_1^3 + x_2^3 \\
\text{subject to } x_1 + x_2 - 1 = 0.
\]

a. Find an optimal solution to the problem.

b. Consider the following penalty problem:

\[
\text{Minimize } x_1^3 + x_2^3 + \mu(x_1 + x_2 - 1)^2.
\]

For each \(\mu > 0\), verify that the optimal solution is unbounded.

c. Note that the optimal solution in parts a and b have different objective values so that the conclusion of Theorem 9.2.2 does not hold true. Explain.

d. Add the constraints \(|x_1| \leq 1\) and \(|x_2| \leq 1\) to the problem, and let \(X = \{(x_1, x_2) : |x_1| \leq 1, |x_2| \leq 1\}\). The penalty problem becomes:

\[
\text{Minimize } x_1^3 + x_2^3 + \mu(x_1 + x_2 - 1)^2 \\
\text{subject to } |x_1| \leq 1, |x_2| \leq 1.
\]

What is the optimal solution for a given \(\mu > 0\)? What is the limit of the sequence of optima as \(\mu \to \infty\)? Note that with the addition of the set \(X\), the conclusion of Theorem 9.2.2 holds true.

[9.8] A new facility is to be placed such that the sum of its squared distances from four existing facilities is minimized. The four facilities are located at the points \((2, 3), (-3, 2), (3, 4),\) and \((-5, -2)\). If the coordinates of the new facility are \(x_1\) and \(x_2\), suppose that \(x_1\) and \(x_2\) must satisfy the restrictions \(3x_1 + 2x_2 = 6,\) \(x_1 \geq 0,\) and \(x_2 \geq 0\).

a. Formulate the problem.

b. Show that the objective function is convex.

c. Find an optimal solution by making use of the KKT conditions.

d. Solve the problem by a penalty function method using a suitable unconstrained optimization technique.

[9.9] The exterior penalty problem can be reformulated as follows: Find \(\sup_{\mu \geq 0} \inf_{x \in X}\{f(x) + \mu \alpha(x)\}\), where \(\alpha\) is a suitable penalty function.
a. Show that the primal problem is equivalent to finding \( \inf_{x \in X} \sup_{\mu \geq 0} \{ f(x) + \mu \alpha(x) \} \). From this, note that the primal and penalty problems can be interpreted as a pair of min-max dual problems.

b. In Theorem 9.2.2 it was shown that
\[
\inf_{x \in X} \sup_{\mu \geq 0} \{ f(x) + \mu \alpha(x) \} = \sup_{\mu \geq 0} \inf_{x \in X} \{ f(x) + \mu \alpha(x) \}
\]
without any convexity assumptions regarding \( f \) or \( \alpha \). For the Lagrangian dual problem of Chapter 6, however, suitable convexity assumptions had to be made to guarantee equality of the optimal objectives of the primal and dual problems. Comment, relating your discussion to Figure 9.3.

[9.10] Consider the problem given by Equation (9.14), and suppose that this problem has a local minimum at \( x^* \), where \( x^* \) is a regular point having a unique associated Lagrange multiplier \( \lambda^* \), and that the Hessian of the Lagrangian \( f(x) + \mu \sum_{i=1}^{\ell} h_i^2(x) + \sum_{i=1}^{\ell} \lambda_i \phi_i(x) \) with respect to \( x \) is positive definite at \( x = x^* \). Define the Lagrangian dual function \( \theta(v) = \min_x \{ f(x) + \mu \sum_{i=1}^{\ell} h_i^2(x) + \sum_{i=1}^{\ell} \lambda_i \phi_i(x) : x \text{ lies in a sufficiently small neighborhood of } x^* \} \). Show that for each \( v \) in some neighborhood of \( v^* \), there exists an unique \( x(v) \) that evaluates \( \theta(v) \) and, moreover, \( x(v) \) is a continuously differentiable function of \( v \).

[9.11] Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) and \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \). For a given value of \( \mu > 0 \), provide an interpretation for the problem to \( \min_x \{ f(x) + \mu \sum_{i=1}^{m} \max\{0, g_i(x)\} + \sum_{i=1}^{\ell} h_i^2(x) \} \) in terms of the problem to \( \min_x \{ \mu \| \epsilon \|^2 + v(\epsilon) \} \), where \( v \) is the perturbation function given by Equation (6.9).

[9.12] Let \( X = \{ x : g_i(x) \leq 0 \text{ for } i = m + 1, \ldots, m + M, \text{ and } h_i(x) = 0 \text{ for } i = \ell + 1, \ldots, \ell + L \} \). Following the derivation of the KKT Lagrange multipliers at optimality when using the exterior penalty function as in Section 9.2, show how and under what conditions one can recover all the Lagrange multipliers at optimality for the problem to minimize \( f(x) \), subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \), \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \), and \( x \in X \).

[9.13] Consider Problem P to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) and \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \). Let \( F_E(x) \) be the exact absolute value penalty function defined by (9.8). Show that if \( \bar{x} \) satisfies the second-order sufficiency conditions for a local minimum of P as stated in Theorem 4.4.2, then for \( \mu \) at least as large as that given by Theorem 9.3.1, \( \bar{x} \) is also a local minimizer for \( F_E \).
[9.14] Consider Problem P to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, ..., m \) and \( h_i(x) = 0 \) for \( i = 1, ..., \ell \). Given \((\bar{u}, \bar{v})\), consider the ALAG inner minimization problem (9.26), and suppose that \( x_k \) is a minimizing solution, so that \( \nabla_x F_{\text{ALAG}}(x_k, \bar{u}, \bar{v}) = 0 \). Show that by requiring \((\bar{u}_{\text{new}}, \bar{v}_{\text{new}})\) to satisfy \( \nabla_x L(x_k, \bar{u}_{\text{new}}, \bar{v}_{\text{new}}) = 0 \) gives \((\bar{u}_{\text{new}}, \bar{v}_{\text{new}})\) as defined by (9.19) and (9.27), where \( L(x, u, v) \) is the usual Lagrangian function for Problem P.

[9.15] Solve the following problem using the method of multipliers, starting with the Lagrange multipliers \( v_1 = v_2 = 0 \) and with the penalty parameters \( \mu_1 = \mu_2 = 1.0:\)

Minimize \( 3x_1 + 2x_2 - 2x_3 \) 
subject to \( x_1^2 + x_2^2 + x_3^2 = 16 \) 
\( 2x_1 - 2x_2 + x_3 = 1. \)

[9.16] Let P: Minimize \( \{f(x) : g_i(x) \leq 0 \text{ for } i = 1, ..., m, \text{ and } h_i(x) = 0 \text{ for } i = 1, ..., \ell \} \), and define P': Minimize \( \{f(x) : g_i(x) + s_i^2 = 0 \text{ for } i = 1, ..., m, \text{ and } h_i(x) = 0 \text{ for } i = 1, ..., \ell \} \). Let \( \bar{x} \) be a KKT point for P with Lagrangian multipliers \( \bar{u} \) and \( \bar{v} \) associated with the inequality and the equality constraints, respectively, such that the strict complementary slackness condition holds true, namely, that \( \bar{u}_i g_i(\bar{x}) = 0 \) for all \( i = 1, ..., m \) and \( \bar{u}_i > 0 \) for each \( i \in I(\bar{x}) = \{i : g_i(\bar{x}) = 0\} \). Furthermore, suppose that the second-order sufficiency conditions of Theorem 4.4.2 hold true for Problem P at this point. Write the KKT conditions and the second-order sufficiency conditions for Problem P' and verify that these conditions are satisfied at \((x, s, \bar{u}, \bar{v})\), where \( s_i^2 = -g_i(\bar{x}) \) for all \( i = 1, ..., m \). Indicate the significance of the strict complementary slackness assumption.

[9.17] Consider the problem to minimize \( f(x) \) subject to \( h_i(x) = 0 \) for \( i = 1, ..., \ell \), where \( f \) and \( h_i \), \( i = 1, ..., \ell \), are continuously twice differentiable. Given any \( \varepsilon = (\varepsilon_1, ..., \varepsilon_\ell) \), define the perturbed problem \( P(\varepsilon) \): Minimize \( \{f(x) : h_i(x) = \varepsilon_i \text{ for } i = 1, ..., \ell \} \). Suppose that P has a local minimum \( \bar{x} \) that is a regular point and that \( \bar{x} \), along with the corresponding unique Lagrange multiplier \( \bar{v} \), satisfies the second-order sufficiency conditions for a strict local minimum. Then show that for each \( \varepsilon \) in a neighborhood of \( \varepsilon = 0 \), there exists a solution \( x(\varepsilon) \) such that (i) \( x(\varepsilon) \) is a local minimum for \( P(\varepsilon) \); (ii) \( x(\varepsilon) \) is a continuous function of \( \varepsilon \), with \( x(0) = \bar{x} \); and (iii) \( \partial f[x(\varepsilon)]/\partial \varepsilon_i |_{\varepsilon = 0} = \nabla_x f(\bar{x})^T \nabla_{\varepsilon_i} x(0) = -\bar{v}_i \) for \( i = 1, ..., \ell \). (Hint: Use the regularity of \( \bar{x} \) and the second-order sufficiency conditions to show the existence of \( x(\varepsilon) \). By the chain rule, the required partial derivative then equals.
\( \nabla f(\bar{x})' \nabla x(0) \), which in turn equals \(-\sum_j \bar{v}_j \nabla h_j(\bar{x})' \nabla x(0) \) from the KKT conditions. Now, use \( h(x(\varepsilon)) = \varepsilon \) to complete the derivation.)

[9.18] Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad (x_1 - 5)^2 + (x_2 - 3)^2 \\
\text{subject to} & \quad 3x_1 + 2x_2 \leq 6 \\
& \quad -4x_1 + 2x_2 \leq 4.
\end{align*}
\]

Formulate a suitable barrier problem with the initial parameter equal to 1. Use an unconstrained optimization technique starting with the point \((0, 0)\) to solve the barrier problem. Provide estimates for the Lagrange multipliers.

[9.19] Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) and \( x \in X \), where \( X = \{ x : g_i(x) \leq 0 \text{ for } i = m + 1, \ldots, m + M, \text{ and } h_j(x) = 0 \text{ for } i = 1, \ldots, \ell \} \). Extend the derivation of Lagrange multipliers at optimality for the barrier function approach, as discussed in Section 9.4, to the case where \( X \) is as defined above in lieu of \( X = \mathbb{R}^n \).

[9.20] By replacing an equality constraint \( h_i(x) = 0 \) by one of the following forms, where \( \varepsilon > 0 \) is a small scalar,

a. \( h_i^2(x) \leq \varepsilon \),
b. \( |h_i(x)| \leq \varepsilon \),
c. \( h_i(x) \leq \varepsilon, -h_i(x) \leq \varepsilon \),

barrier functions could be used to handle equality constraints. Discuss in detail the implications of these formulations. Using this approach, solve the following problem with \( \varepsilon = 0.05 \):

\[
\begin{align*}
\text{Minimize} & \quad 2x_1^2 + x_2^2 \\
\text{subject to} & \quad 3x_1 + 2x_2 = 6.
\end{align*}
\]

[9.21] This exercise describes several strategies for modifying the barrier parameter \( \mu \). Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad 2(x_1 - 3)^2 + (x_2 - 5)^2 \\
\text{subject to} & \quad 2x_1^2 - x_2 \leq 0.
\end{align*}
\]

Using the auxiliary function \( 2(x_1 - 3)^2 + (x_2 - 5)^2 - \mu[1/(2x_1^2 - x_2)] \) and adopting the cyclic coordinate method, solve the above problem starting from the point \( x_1 = (0, 10)' \) under the following strategies for modifying \( \mu \).
a. Starting from $x_1$, solve the barrier problem for $\mu_1 = 10.0$ resulting in $x_2$. Then starting from $x_2$, solve the problem with $\mu_2 = 0.01$.

b. Starting from the point $(0, 10)$, solve the barrier problem for $\mu_1 = 0.01$.

c. Apply the algorithm described in Section 9.3 by using the successively decreasing values of $\mu = 10.0, 1.00, 0.10, \text{ and } 0.01$.

d. Starting from $x_1$, solve the barrier problem for $\mu_1 = 0.001$.

Which of the above strategies would you recommend, and why? Also, in each case above, derive an estimate for the Lagrange multiplier associated with the single constraint.

[9.22] Compare the different exterior penalty and the interior barrier function methods in detail. Emphasize the advantages and disadvantages of both methods.

[9.23] In the methods discussed in this chapter, the same penalty or barrier parameter was often used for all the constraints. Do you see any advantages for using different parameters for the various constraints? Propose some schemes for updating these parameters. How can you modify Theorems 9.2.2 and 9.4.3 to handle this situation?

[9.24] To use a barrier function method, we must find a point $x \in X$ having $g_i(x) < 0$ for $i = 1, \ldots, m$. The following procedure is suggested for obtaining such a point.

Initialization Step Select $x_1 \in X$, let $k = 1$, and go to the Main Step.

Main Step

1. Let $I = \{i : g_i(x_k) < 0\}$. If $I = \{1, \ldots, m\}$, stop with $x_k$ satisfying $g_i(x_k) < 0$ for all $i$. Otherwise, select $j \in I$ and go to Step 2.

2. Use a barrier function method to solve the following problem starting with $x_k$:

   
   \[
   \begin{align*}
   \text{Minimize} & \quad g_i(x) \\
   \text{subject to} & \quad g_i(x) < 0 \text{ for } i \in I \\
   & \quad x \in X.
   \end{align*}
   \]

Let $x_{k+1}$ be an optimal solution. If $g_i(x_{k+1}) \geq 0$, stop; the set $\{x \in X : g_i(x) < 0 \text{ for } i = 1, \ldots, m\}$ is empty. Otherwise, replace $k$ by $k + 1$ and go to Step 1.

a. Show that the above approach stops in, at most, $m$ iterations with a point $x \in X$ satisfying $g_i(x) < 0$ for $i = 1, \ldots, m$ or with the conclusion that no such point exists.

b. Using the above approach, find a point satisfying $2x_1 + x_2 < 2$ and $2x_1^2 - x_2 < 0$, starting from the point $(2, 0)$. 

[9.25] Consider the problem to minimize \( f(x) \) subject to \( x \in X, g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) and \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \). A mixed penalty-barrier auxiliary function is of the form \( f(x) + \mu B(x) + (1/\mu)\alpha(x) \), where \( B \) is a barrier function that handles the inequality constraints and \( \alpha \) is a penalty function that handles the equality constraints. The following result generalizes Theorems 9.2.2 and 9.4.3:

\[
\inf\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\} = \lim_{\mu \to 0^+} \sigma(\mu)
\]

\[
\mu B(x_\mu) \to 0 \quad \text{and} \quad \frac{1}{\mu} \alpha(x_\mu) \to 0 \quad \text{as} \quad \mu \to 0^+,
\]

where

\[
\sigma(\mu) = \inf\left\{ f(x) + \mu B(x) + \frac{1}{\mu} \alpha(x) : x \in X, g(x) < 0 \right\}
\]

\[
= f(x_\mu) + \mu B(x_\mu) + \frac{1}{\mu} \alpha(x_\mu).
\]

a. Prove the above result after making appropriate assumptions.

b. State a precise algorithm for solving a nonlinear programming problem using a mixed penalty-barrier function approach, and illustrate by solving the following problem:

Minimize \( 3e^{x_1} - 2x_1x_2 + 2x_2^2 \)

subject to \( x_1^2 + x_2^2 = 9 \)
\( 3x_1 + 2x_2 \leq 6. \)

c. Discuss the possibility of using two parameters \( \mu_1 \) and \( \mu_2 \) so that the mixed penalty-barrier auxiliary function is of the form \( f(x) + \mu_1 B(x) + (1/\mu_2)\alpha(x) \). Using this approach, solve the following problem by a suitable unconstrained optimization technique, starting from the point \((0, 0)\) and initially letting \( \mu_1 = 1.0 \) and \( \mu_2 = 2.0 \):

Maximize \( -2x_1^2 + 2x_1x_2 + 3x_2^2 - e^{-x_1-x_2} \)

subject to \( x_1^2 + x_2^2 - 9 = 0 \)
\( 3x_1 + 2x_2 \leq 6. \)

(The method described in this exercise is credited to Fiacco and McCormick [1968].)
[9.26] In this exercise we describe a parameter-free penalty function method for solving a problem of the form to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$ and $h_i(x) = 0$ for $i = 1, \ldots, \ell$.

**Initialization Step** Choose a scalar $L_1 < \inf\{f(x) : g_i(x) \leq 0 \text{ for } i = 1, \ldots, m, \text{ and } h_i(x) = 0 \text{ for } i = 1, \ldots, \ell\}$. Let $k = 1$, and go to the Main Step.

**Main Step** Solve the following problem:

\[
\text{Minimize } \beta(x) \\
\text{subject to } x \in \mathbb{R}^n,
\]

where

\[
\beta(x) = \left[\max\{0, f(x) - L_k\}\right]^2 + \sum_{i=1}^{m} \left[\max\{0, g_i(x)\}\right]^2 + \sum_{i=1}^{\ell} \left|h_i(x)\right|^2.
\]

Let $x_k$ be an optimal solution, let $L_{k+1} = f(x_k)$, replace $k$ by $k + 1$, and repeat the Main Step.

a. Solve the following problem by the approach above, starting with $L_1 = 0$ and initiating the optimization process from the point $x = (0, -3)^t$:

\[
\text{Minimize } 2(x_1 - 3)^2 + (x_2 - 5)^2 \\
\text{subject to } 2x_1^2 - x_2 \leq 0.
\]

b. Compare the trajectory of points generated in Part a with the points generated in Exercise 9.3.

c. At iteration $k$, if $x_k$ is feasible to the original problem, show that it must be optimal.

d. State the assumptions under which the above method converges to an optimal solution, and prove convergence.

[9.27] In this exercise we describe a parameter-free barrier function method for solving a problem of the form to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$.

**Initialization Step** Choose $x_1$ such that $g_i(x_1) < 0$ for $i = 1, \ldots, m$. Let $k = 1$, and go to the Main Step.

**Main Step** Let $X_k = \{x : f(x) - f(x_k) < 0, g_i(x) < 0 \text{ for } i = 1, \ldots, m\}$. Let $x_{k+1}$ be an optimal solution to the following problem:

\[
\text{Minimize } \frac{-1}{f(x) - f(x_k)} - \sum_{i=1}^{m} \frac{1}{g_i(x)} \\
\text{subject to } x \in X_k.
\]
Replace \( k \) by \( k + 1 \), and repeat the Main Step. (The constraint \( x \in X_k \) could be handled implicitly provided that we started the optimization process from \( x = x_k \).)

a. Solve the following problem by the approach above, starting from \( x_1 = (0, 10)^t \).

Minimize \( 2(x_1 - 3)^2 + (x_2 - 5)^2 \)
subject to \( 2x_1^2 - x_2 \leq 0 \).

b. Compare the trajectory of points generated in Part a with the points generated in Exercise 9.21.

c. State the assumptions under which the above method converges to an optimal solution, and prove convergence.

[9.28] Consider the following problem:

Minimize \( f(x) \)
subject to \( h(x) = 0 \),

where \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) and \( h: \mathbb{R}^n \rightarrow \mathbb{R}^\ell \) are differentiable. Let \( \mu > 0 \) be a large penalty parameter, and consider the penalty problem to minimize \( q(x) \) subject to \( x \in \mathbb{R}^n \), where \( q(x) = f(x) + \mu \sum_{i=1}^\ell h_i^2(x) \). The following method is suggested to solve the penalty problem (see Luenberger [1984] for a detailed motivation of the method).

**Initialization Step** Choose a point \( x_1 \in \mathbb{R}^n \), let \( k = 1 \), and go to the Main Step.

**Main Step**

1. Let the \( \ell \times n \) matrix \( \nabla h(x_k) \) denote the Jacobian of \( h \) at \( x_k \). Let \( A = BB \), where \( B = \nabla h(x_k) \nabla h(x_k)^t \). Let \( d_k \) be as given below and go to Step 2.

\[
d_k = -\frac{1}{2\mu} \nabla h(x_k)^t B^{-1} \nabla h(x_k) \nabla q(x_k) .
\]

2. Let \( \lambda_k \) be an optimal solution to the problem to minimize \( q(x_k + \lambda d_k) \) subject to \( \lambda \in \mathbb{R} \), and let \( w_k = x_k + \lambda_k d_k \). Go to Step 3.

3. Let \( \alpha_k = -\nabla q(w_k) \), and let \( \alpha_k \) be an optimal solution to the problem to minimize \( q(w_k + \alpha d_k) \) subject to \( \alpha \in \mathbb{R} \). Let \( x_{k+1} = w_k + \alpha_k d_k \), replace \( k \) by \( k + 1 \), and go to Step 1.

a. Apply the above method to solve the following problem by letting \( \mu = 100 \):
Minimize \( 2x_1^2 + 2x_1x_2 + 3x_2^2 - 2x_1 + 3x_2 \)
subject to \( 3x_1 + 2x_2 = 6. \)

b. The above algorithm could easily be modified to solve a problem having equality and inequality constraints. In this case, let the penalty problem be to minimize \( q(x) \), where \( q(x) = f(x) + \mu \sum_{i=1}^{\ell} h_i^2(x) + \mu \sum_{i=1}^{m} [\max\{0, g_i(x)\}]^2 \). In the description of the algorithm, \( h(x_k) \) is replaced by \( F(x_k) \), where \( F(x_k) \) consists of the equality constraints and the inequality constraints that are either active or violated at \( x_k \). Use this modified procedure to solve the following problem, by letting \( \mu = 100. \)

Minimize \( 2x_1^2 + 2x_1x_2 + 3x_2^2 - 2x_1 + 3x_2 \)
subject to \( 3x_1 + 2x_2 \geq 6 \)
\( x_1, x_2 \geq 0. \)

[9.29] Consider the problem to minimize \( f(x) \) subject to \( h_i(x) = 0 \) for \( i = 1, \ldots, \ell \), and suppose that it has an optimal solution \( \bar{x} \). The following procedure for solving the problem is credited to Morrison [1968].

**Initialization Step** Choose \( L_1 \leq f(\bar{x}) \), let \( k = 1 \), and go to the Main Step.

**Main Step**

1. Let \( x_k \) be an optimal solution for the problem to minimize \( [f(x) - L_k]^2 + \sum_{i=1}^{\ell} h_i^2(x) \). If \( h_i(x_k) = 0 \) for \( i = 1, \ldots, \ell \), stop with \( x_k \) as an optimal solution to the original problem; otherwise, go to Step 2.

2. Let \( L_{k+1} = L_k + v^{1/2} \), where \( v = [f(x_k) - L_k]^2 + \sum_{i=1}^{\ell} h_i^2(x_k) \).
Replace \( k \) by \( k + 1 \) and go to Step 1.

a. Show that if \( h_i(x_k) = 0 \) for \( i = 1, \ldots, \ell \), then \( L_{k+1} = f(x_k) = f(\bar{x}) \), and \( x_k \) is an optimal solution to the original problem.

b. Show that \( f(x_k) \leq f(\bar{x}) \) for each \( k \).

c. Show that \( L_k \leq f(\bar{x}) \) for each \( k \) and that \( L_k \to f(\bar{x}) \).

d. Using the above method, solve the following problem:

Minimize \( 2x_1 + 3x_2 - 2x_3 \)
subject to \( x_1^2 + x_2^2 + x_3^2 = 16 \)
\( 2x_1 - 2x_2^2 + x_3 = 1. \)
[9.30] Consider the problem to minimize $2x_1 + 3x_2 + x_3$, subject to $3x_1 + 2x_2 + 4x_3 = 9$, $x \geq 0$. Solve this problem using the primal-dual path-following method using a starting solution (at iteration $k = 0$) of $x_k = (1, 1, 1)^t$, $v_k = -1$, and $\mu_k = 5$, where $v$ is the dual variable associated with the single equality constraint. Is this starting solution on the central path?

[9.31] Re-solve the example of Exercise 9.30 using the predictor-corrector path-following algorithm described in Section 9.5, starting with the same solution as that given in Exercise 9.30 and adopting the following rule to update the parameter $\mu$:

$$
\mu_{k+1} = \left( \frac{u_{k+1}'x_{k+1}}{nu_k'x_k} \right)^2.
$$

Give an interpretation of this formula.

[Hint: Examine Equations (9.40) and (9.41) and observe that

$$
\mu_{k+1} = \left( \frac{u_{k+1}'x_{k+1}}{u_k'x_k} \right)^2 \frac{u_k'x_k}{n}.
$$

[9.32] Consider the problem of Example 9.5.1. Using (9.43), obtain a simplified closed-form expression for $\tilde{w} = (\tilde{x}, \tilde{u}, \tilde{v})$ in terms of $\tilde{w} = (\bar{x}, \bar{u}, \bar{v})$ and $\tilde{\mu}$. Hence, obtain the sequence of iterates generated by the primal-dual path-following algorithm starting with $\bar{x} = (2/9, 8/9)^t$, $\bar{u} = (4, 1)^t$, $\bar{v} = -1$, and $\bar{\mu} = 8/9$, using $\theta = \delta = 0.35$.

[9.33] Consider the linear programming Problem P to minimize $c'x$ subject to $Ax = 0$, $e'x = 1$, and $x \geq 0$, where $A$ is an $m \times n$ matrix of rank $m < n$ and $e$ is a vector of $n$ ones. For a given $\mu > 0$, define $P_\mu$ to be the problem to minimize $c'x + \mu \sum_{j=1}^n x_j \ln(x_j)$ subject to $Ax = 0$, and $e'x = 1$, with $x > 0$ treated implicitly.

a. Show that any linear programming problem of the type to minimize $c'x$ subject to $Ax = b$, $x \geq 0$ that has a bounded feasible region can be transformed into the form of Problem P. (This form of linear program is due to Karmarkar [1984].)

b. What happens to the negative entropy function $\mu \sum_{j=1}^n x_j \ln(x_j)$ as some $x_j \rightarrow 0^+$? How does this compare with the logarithmic barrier function (9.29b)? Examine the partial derivatives of this negative entropy function with respect to $x_j$ as $x_j \rightarrow 0^+$ and, hence, justify
why this function might act as a barrier. Hence, show that the optimal solution to $P_\mu$ approaches the optimum to $P$ as $\mu \to 0^+$. (This result is due to Fang [1990].)

c. Consider Problem $P$ to minimize $-x_3$ subject to $x_1 - x_2 = 0$, $x_1 + x_2 + x_3 = 1$, and $x \geq 0$. For the corresponding Problem $P_\mu$, construct the Lagrangian dual to maximize $\theta(\pi)$, $\pi \in \mathbb{R}$, where $\pi$ is the Lagrange multiplier associated with the constraint $x_1 - x_2 = 0$. Using the KKT conditions, show that this is equivalent to minimizing

$$\ln[e^{(\pi/\mu)-1} + e^{(-\pi/\mu)-1} + e^{(1/\mu)-1}]$$

and hence that $\pi = 0$ is optimal. Accordingly, show that $P_\mu$ is solved by $x_1 = x_2 = 1/(2 + e^{1/\mu})$ and $x_3 = e^{1/\mu}/(2 + e^{1/\mu})$ and that the limit of this solution as $\mu \to 0^+$ solves $P$.

[9.34] Consider the Problem $P$ to minimize $c^T x$ subject to $Ax = b$ and $x \geq 0$, where $A$ is an $m \times n$ matrix of rank $m < n$. Suppose that we are given a feasible solution $\bar{x} > 0$. For a barrier parameter $\mu > 0$, consider Problem $BP$ to minimize

$$f(x) = c^T x - \mu \sum_{j=1}^n \ln(x_j)$$

subject to $Ax = b$, with $x > 0$ treated implicitly.

a. Find the direction $d$ that minimizes the second-order Taylor series approximation to $f(\bar{x} + d)$ subject to $Ad = 0$. Interpret this as a projected Newton direction.

b. Consider the problem $\delta(\bar{x}, \mu) = \min_{(u,v)} [(1/\mu)\bar{X}u - e]$ subject to $A^Tv + u = c$. Find the optimum solution $(\bar{u}, \bar{v})$. Interpret this in relation to the dual to $P$ and Equation (9.39). Show that the direction of Part a satisfies $d = \bar{x} - (1/\mu)\bar{X}^2\bar{u}$.

c. Consider the following algorithm due to Roos and Vial [1988]. Start with some $\mu > 0$ and a solution $\bar{x} > 0$ such that $A\bar{x} = b$ and $\delta(\bar{x}, \mu) \leq 1/2$. Let $\theta = 1/(6\sqrt{n})$. Find $d$ as in Part b and revise $\bar{x}$ to $\bar{x} + d$. While the duality gap is not small enough, repeat this main step by revising $\mu$ to $\mu(1 - \theta)$, computing $d = \bar{x} - (1/\mu)\bar{X}^2\bar{u}$ as given by Part b, and setting the revised solution $\bar{x}$ to $\bar{x} + d$. Illustrate this algorithm for the problem of Example 9.5.1.

[9.35] Consider the problem of Example 9.5.1. Construct the artificial primal-dual pair of linear Programs $P'$ and $D'$ given in (9.57). Using the starting solution (9.58), perform at least two iterations of the primal-dual path-following algorithm.
Notes and References

The use of penalty functions to solve constrained problems is generally attributed to Courant. Subsequently, Camp [1955] and Pietrzykowski [1962] discussed this approach to solve nonlinear problems. The latter reference also gives a convergence proof. However, significant progress in solving practical problems by the use of penalty methods follows the classic work of Fiacco and McCormick under the title SUMT (sequential unconstrained minimization technique). The interested reader may refer to Fiacco and McCormick [1964a,b, 1966, 1967b, 1968] and Zangwill [1967c, 1969]. See Himmelblau [1972b], Lootsma [1968a,b], and Osborne and Ryan [1970] for the performance of different penalty functions on test problems.

Luenberger [1973a/1984] discusses the use of a generalized quadratic penalty function of the form \( \mu h(x) \Gamma h(x) \), where \( \Gamma \) is a symmetric \( \ell \times \ell \) positive definite matrix. He also discusses a combination of penalty function and gradient projection methods (see Exercise 9.28). We also refer the reader to Luenberger [1973a/1984] for further details on eigenvalue analyses of the Hessians of the various penalty functions and their effect on the convergence characteristics of penalty-based algorithms. Best et al. [1981] discuss how Newton’s method can be used to efficiently purify an approximate solution determined by penalty methods to a more accurate optimum. For the use of penalty functions in conducting sensitivity analyses in nonlinear programming, see Fiacco [1983].

The barrier function approach was first proposed by Carroll [1961] under the name created response surface technique. The approach was used to solve nonlinear inequality constrained problems by Box et al. [1969] and Kowalik [1966]. The barrier function approach has been investigated thoroughly and popularized by Fiacco and McCormick [1964a,b, 1968]. In Exercise 9.25 we introduce the mixed penalty-barrier auxiliary functions studied by Fiacco and McCormick [1968]. Here, equality and inequality constraints are handled, respectively, by a penalty term and a barrier term. Also see Belmore et al. [1970], Greenberg [1973b], and Raghavendra and Rao [1973].

The numerical problem of how to change the parameters of the penalty and barrier functions have been investigated by several authors. See Fiacco and McCormick [1968] and Himmelblau [1972b] for a detailed discussion. These references also give the computational experience for numerous problems. Bazaraa [1975], Lasdon [1972], and Lasdon and Ratner [1973] discuss effective unconstrained optimization algorithms for solving penalty and barrier functions. For eigenvalue analyses relating to the convergence behavior of algorithms applied to the barrier function problem, see Luenberger [1973a/1984].

Several extensions to the concepts of penalty and barrier functions have been made. First, to avoid the difficulties associated with ill-conditioning as the penalty parameter approaches infinity and as the barrier parameter approaches zero, several parameter-free methods have been proposed. This concept was introduced in Exercises 9.26 and 9.27. For further details of this topic, see Fiacco and McCormick [1968], Huard’s method of centers [1967], and Lootsma...
Another popular variant that has good computational characteristics is the \textit{shifted/modified barrier method} (see Polak, 1992).

Through the absolute value $\ell_1$ penalty function, we introduced the concept of exact penalty functions in which a single unconstrained minimization problem, with a reasonably sized penalty parameter, can yield an optimum solution to the original problem. This was first introduced by Pietrzykowski [1969] and Fletcher [1970b], and has been studied by Bazarraa and Goode [1982], Coleman and Conn [1982a,b], Conn [1985], Evans et al. [1973], Fletcher [1973, 1981b, 1985], Gill et al. [1981], Han [1979], and Mayne [1980].

A popular and useful exact penalty approach that uses both a Lagrangian multiplier term and a penalty term in the auxiliary function is the \textit{method of multipliers} or \textit{augmented Lagrangian} (ALAG) approach. This approach was proposed independently by Hestenes [1969] and Powell [1969]. Again, the motivation is to avoid the ill-conditioning difficulties encountered by the classical approach as the penalty parameter approaches infinity. For further details, refer to Bertsekas [1975a,c,d, 1976a,b], Boggs and Tolle [1980], Fletcher [1975, 1987], Hestenes [1980b], Miele et al. [1971a,b], Pierre and Lowe [1975], Rockafellar [1973a,b, 1974], and Tapia [1977]. Conn et al. [1988a] discuss a method for effectively incorporating simple bounds within the constraint set. Fletcher [1985, 1987] also suggests that in a mix of nonlinear and linear constraints, it might be worthwhile to incorporate only the nonlinear constraints in the penalty function. Sen and Sherali [1986b] and Sherali and Ulular [1989] discuss a primal-dual conjugate subgradient algorithm for solving differentiable or nondifferentiable, decomposable problems using ALAG penalty functions and Lagrangian dual functions in concert with each other. Polyak and Tret’iakov [1972] and Delbos and Gilbert [2004] discuss ALAG approaches for linear and quadratic programming problems.

In Section 9.5 we have presented a polynomial-time \textit{primal-dual path-following} algorithm as developed by Monteiro and Adler [1989a] based on Frisch’s [1955] logarithmic barrier function and motivated by Karmarkar’s [1984] polynomial-time algorithm for linear programming. This algorithm also readily extends to solve convex quadratic programs with the same order of complexity; see Monteiro and Adler [1989b]. The concept of “good” or polynomially bounded algorithms was proposed independently by Edmonds [1965] and Cobham [1965]. See Cook [1971], Karp [1972], Garey and Johnson [1979], and Papadimitriou and Steiglitz [1982] for further reading on this subject. For a discussion on complexity issues in linear programming and purification schemes, we refer the reader to Bazarraa et al. [2005] and Murty [1983]. For an extension of this algorithmic concept to one that employs both first- and higher-order power series approximations to a weighted barrier path, see Monteiro et al. [1990]. Computational aspects and implementation details of this class of algorithms is discussed in Choi et al. [1990], McShane et al. [1989], Lustig et al. [1990, 1994a,b], and Mehrotra [1990]; and many useful ideas, such as extrapolation of iterates generated to accelerate convergence, can be traced to Fiacco and McCormick [1968]. For superlinearly convergent polynomial time primal-dual path-following methods, see Mehrotra [1993], Tapia et al. [1995], Ye et al. [1993], Zhang et al. [1992], and Zhang and Tapia [1993]. Alternative
Path-following algorithms are described, for example, by Ben Daya and Shetty [1988], Gonzaga [1987], Peng et al. [2002], Renegar [1988], Roos and Vial [1988], Vaidya [1987], and Ye and Todd [1987], among others, and are motivated by Sonnevend’s [1985] method of centers and Megiddo’s [1986] trajectory to optimality. A non-polynomial-time algorithm based on using Newton’s method with the inverse barrier function is discussed by den Hertog et al. [1991]. The popular predictor-corrector variants of primal-dual path-following algorithms were introduced by Mehrotra [1991, 1992], and implemented computationally by Lustig and Li [1992]. Kojima et al. [1993] (see implementation aspects in Lustig et al. [1994a,b]) and Zhang and Zhang [1995] provide convergence analyses and polynomial complexity proofs for such variants. Carpenter et al. [1993] explore higher-order variants of predictor-corrector methods. For extensions of interior point methods to quadratic and convex nonlinear programs, see the exposition given in Hertog [1994] and Nesterov and Nemirovskii [1993]. For a combination of $\ell_1$ penalty and interior point methods for nonlinear optimization, see Gould et al. [2003]. See Todd [1989], Terlaky [1998], and Martin [1999] for surveys of other variants of Karmarkar’s algorithm.
Chapter Methods of Feasible Directions

The class of feasible direction methods solves a nonlinear programming problem by moving from a feasible point to an improved feasible point. The following strategy is typical of feasible direction algorithms. Given a feasible point $x_k$, a direction $d_k$ is determined such that for $\lambda > 0$ and sufficiently small, the following two properties are true: (1) $x_k + \lambda d_k$ is feasible, and (2) the objective value at $x_k + \lambda d_k$ is better than the objective value at $x_k$. After such a direction is determined, a one-dimensional optimization problem is solved to determine how far to proceed along $d_k$. This leads to a new point $x_{k+1}$, and the process is repeated. Since primal feasibility is maintained during this optimization process, these procedures are often referred to as primal methods. Methods of this type are frequently shown to converge to KKT solutions or, sometimes, to FJ points. We ask the reader to review Chapter 4 to assess the worth of such solutions.

Following is an outline of the chapter.

Section 10.1: Method of Zoutendijk
In this section we show how to generate an improving feasible direction by solving a subproblem that is usually a linear program. Problems having linear constraints and problems having nonlinear constraints are both considered.

Section 10.2: Convergence Analysis of the Method of Zoutendijk
We show that the algorithmic map of Section 10.1 is not closed, so that convergence is not guaranteed. A modification of the basic algorithm credited to Topkis and Veinott [1967] that guarantees convergence is presented.

Section 10.3: Successive Linear Programming Approach
We describe a penalty-based successive linear programming approach that combines the ideas of sequentially using a linearized feasible direction subproblem along with $\ell_1$ penalty function concepts to yield an efficient and convergent algorithm. For vertex optimal solutions, a quadratic convergence rate is possible but otherwise, convergence can be slow.

Section 10.4: Successive Quadratic Programming or Projected Lagrangian Approach
To obtain a quadratic or superlinear convergence behavior, even for nonvertex solutions, we can adopt Newton’s method or quasi-Newton methods to solve the KKT optimality conditions. This leads to a
successive quadratic programming or projected Lagrangian approach in which the feasible direction subproblem is a quadratic problem, with the Hessian of the objective function being that of the Lagrangian function and the constraints representing first-order approximations. We describe both a rudimentary version of this method and a globally convergent variant using the $\ell_1$ penalty function as either a merit function or, more actively, using it in the subproblem objective function itself. The associated Maratos effect is also discussed.

Section 10.5: Gradient Projection Method of Rosen In this section we describe how to generate an improving feasible direction for a problem having linear constraints by projecting the gradient of the objective function onto the nullspace of the gradients of the binding constraints. A convergent variant is also presented.

Section 10.6: Reduced Gradient Method of Wolfe and the Generalized Reduced Gradient Method The variables are represented in terms of an independent subset of the variables. For a problem having linear constraints, an improving feasible direction is determined based on the gradient vector in the reduced space. A generalized reduced gradient variant for nonlinear constraints is also discussed.

Section 10.7: Convex-Simplex Method of Zangwill We describe the convex-simplex method for solving a nonlinear program in the presence of linear constraints. The method is identical to the reduced gradient method except that an improving feasible direction is determined by modifying only one nonbasic variable and adjusting the basic variables accordingly. If the objective function is linear, then the convex-simplex method reduces to the simplex method of linear programming.

Section 10.8: Effective First- and Second-Order Variants of the Reduced Gradient Method We unify and extend the reduced gradient and convex simplex methods, introducing the concept of suboptimization through the use of superbasic variables. We also discuss the use of second-order functional approximations for finding a direction of movement in the reduced space of the superbasic variables.

10.1 Method of Zoutendijk

In this section we describe the method of feasible directions due to Zoutendijk. At each iteration, the method generates an improving feasible direction and then optimizes along that direction. Definition 10.1.1 reiterates the notion of an improving feasible direction from Chapter 4.

10.1.1 Definition

Consider the problem to minimize $f(x)$ subject to $x \in S$, where $f: \mathbb{R}^n \to \mathbb{R}$ and $S$ is a nonempty set in $\mathbb{R}^n$. A nonzero vector $d$ is called a feasible direction at $x \in$
Methods of Feasible Directions

539

If there exists a $\delta > 0$ such that $x + \lambda d \in S$ for all $\lambda \in (0, \delta)$. Furthermore, $d$ is called an improving feasible direction at $x \in S$ if there exists a $\delta > 0$ such that $f(x + \lambda d) < f(x)$ and $x + \lambda d \in S$ for all $\lambda \in (0, \delta)$.

Case of Linear Constraints

We first consider the case where the feasible region $S$ is defined by a system of linear constraints, so that the problem under consideration is of the form:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad Ax \leq b \\
& \quad Qx = q.
\end{align*}
\]

Here $A$ is an $m \times n$ matrix, $Q$ is an $\ell \times n$ matrix, $b$ is an $m$-vector, and $q$ is an $\ell$-vector. Lemma 10.1.2 gives a suitable characterization of a feasible direction and a sufficient condition for an improving direction. In particular, $d$ is an improving feasible direction if $A_1^T d \leq 0$, $Q d = 0$, and $\nabla f(x)^T d < 0$. The proof of the lemma is straightforward and is left as an exercise to the reader (see Theorem 3.1.2 and Exercise 10.3).

10.1.2 Lemma

Consider the problem to minimize $f(x)$ subject to $Ax \leq b$ and $Qx = q$. Let $x$ be a feasible solution, and suppose that $A_1^T x = b_1$ and $A_2^T x < b_2$, where $A^T$ is decomposed into $(A_1^T, A_2^T)$ and $b^T$ is decomposed into $(b_1^T, b_2^T)$. Then a nonzero vector $d$ is a feasible direction at $x$ if and only if $A_1^T d \leq 0$ and $Q d = 0$. If $\nabla f(x)^T d < 0$, then $d$ is an improving direction.

Geometric Interpretation of Improving Feasible Directions

We now illustrate the set of improving feasible directions geometrically by the following example.

10.1.3 Example

Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad (x_1 - 6)^2 + (x_2 - 2)^2 \\
\text{subject to} & \quad -x_1 + 2x_2 \leq 4 \\
& \quad 3x_1 + 2x_2 \leq 12 \\
& \quad -x_1 \leq 0 \\
& \quad -x_2 \leq 0.
\end{align*}
\]
Let \( x = (2, 3)' \) and note that the first two constraints are binding. In particular, the matrix \( A_1 \) of Lemma 10.1.2 is given by

\[
A_1 = \begin{bmatrix}
-1 & 2 \\
3 & 2
\end{bmatrix}.
\]

Therefore, \( d \) is a feasible direction at \( x \) if and only if \( A_1 d \leq 0 \), that is, if and only if

\[
-d_1 + 2d_2 \leq 0
\]

\[
3d_1 + 2d_2 \leq 0.
\]

The collection of these directions, where the origin is translated to the point \( x \) for convenience, forms the cone of feasible directions shown in Figure 10.1. Note that if we move a short distance starting from \( x \) along any vector \( d \) satisfying the above two inequalities, we would remain in the feasible region.

If a vector \( d \) satisfies \( 0 > \nabla f(x)' d = -8d_1 + 2d_2 \), then \( d \) is an improving direction. Thus, the collection of improving directions is given by the open half-space \( \{(d_1, d_2) : -8d_1 + 2d_2 < 0\} \). The intersection of the cone of feasible directions with this half-space gives the set of all improving feasible directions.

![Figure 10.1 Improving feasible directions.](image-url)
Generating Improving Feasible Directions

Given a feasible point \( x \), as shown in Lemma 10.2.2, a nonzero vector \( d \) is an improving feasible direction if \( \nabla f(x)'d < 0 \), \( A_1d \leq 0 \), and \( Qd = 0 \). A natural method for generating such a direction is to minimize \( \nabla f(x)'d \) subject to the constraints \( A_1d \leq 0 \) and \( Qd = 0 \). Note, however, that if a vector \( \tilde{d} \) such that \( \nabla f(x)'\tilde{d} < 0 \), \( A_1\tilde{d} \leq 0 \), and \( Q\tilde{d} = 0 \) exists, then the optimal objective value of the foregoing problem is \(-\infty\) by considering \( \lambda \tilde{d} \), where \( \lambda \to \infty \). Thus, a constraint that bounds the vector \( d \) or the objective function must be introduced. Such a restriction is usually referred to as a normalization constraint. We present below three problems for generating an improving feasible direction. Each problem uses a different normalization constraint.

**Problem P1**: Minimize \( \nabla f(x)'d \)
subject to
\[
\begin{align*}
A_1d &\leq 0 \\
Qd &= 0 \\
-1 &\leq d_j \leq 1 \text{ for } j = 1, \ldots, n.
\end{align*}
\]

**Problem P2**: Minimize \( \nabla f(x)'d \)
subject to
\[
\begin{align*}
A_1d &\leq 0 \\
Qd &= 0 \\
d'd &\leq 1.
\end{align*}
\]

**Problem P3**: Minimize \( \nabla f(x)'d \)
subject to
\[
\begin{align*}
A_1d &\leq 0 \\
Qd &= 0 \\
\nabla f(x)'d &\geq -1.
\end{align*}
\]

Problems P1 and P3 are linear in the variables \( d_1, \ldots, d_n \) and can be solved by the simplex method. Problem P2 contains a quadratic constraint but could be considerably simplified (see Exercise 10.29). Since \( d = 0 \) is a feasible solution to each of the above problems, and since its objective value is equal to 0, the optimal objective value of Problems P1, P2, and P3 cannot be positive. If the minimal objective function value of Problems P1, P2, and P3 is negative, then by Lemma 10.1.2, an improving feasible direction is generated. On the other hand, if the minimal objective function value is equal to zero, then \( x \) is a KKT point, as shown below.
10.1.4 Lemma

Consider the problem to minimize \( f(x) \) subject to \( Ax \leq b \) and \( Qx = q \). Let \( x \) be a feasible solution such that \( A_1x = b_1 \) and \( A_2x < b_2 \), where \( A' = (A_1', A_2') \) and \( b' = (b_1', b_2') \). Then for each \( i = 1, 2, 3 \), \( x \) is a KKT point if and only if the optimal objective value of Problem \( P_i \) is equal to zero.

Proof

The vector \( x \) is a KKT point if and only if there exist a vector \( u \geq 0 \) and a vector \( v \) such that \( \nabla f(x) + A'u + Q'v = 0 \). By Corollary 3 to Theorem 2.4.5, this system is solvable if and only if the system \( \nabla f(x)'d < 0 \), \( A_1d \leq 0 \), \( Qd = 0 \) has no solution, that is, if and only if the optimal objective value of each of Problems \( P_1 \), \( P_2 \), and \( P_3 \) is equal to zero. This completes the proof.

Line Search

So far, we have seen how to generate an improving feasible direction or conclude that the current vector is a KKT point. Now let \( x_k \) be the current vector, and let \( d_k \) be an improving feasible direction. The next point, \( x_{k+1} \), is given by \( x_k + \lambda_k d_k \), where the step size \( \lambda_k \) is obtained by solving the following one-dimensional problem:

\[
\begin{align*}
\text{Minimize} & \quad f(x_k + \lambda d_k) \\
\text{subject to} & \quad A(x_k + \lambda d_k) \leq b \\
& \quad Q(x_k + \lambda d_k) = q \\
& \quad \lambda \geq 0.
\end{align*}
\]

Now, suppose that \( A' \) is decomposed into \( (A_1', A_2') \) and \( b' \) is decomposed into \( (b_1', b_2') \) such that \( A_1x_k = b_1 \) and \( A_2x_k < b_2 \). Then the above problem could be simplified as follows. First note that \( Qx_k = q \) and \( Qd_k = 0 \), so that the constraint \( Q(x_k + \lambda d_k) = q \) is redundant. Since \( A_1x_k = b_1 \) and \( A_1d_k \leq 0 \), then \( A_1(x_k + \lambda d_k) \leq b_1 \) for all \( \lambda \geq 0 \). Hence, we only need to restrict \( \lambda \) so that \( \lambda A_2d_k \leq b_2 - A_2x_k \). It thus follows that the above problem reduces to the following line search problem, which could be solved by one of the techniques discussed in Sections 8.1 to 8.3:

\[
\begin{align*}
\text{Minimize} & \quad f(x_k + \lambda d_k) \\
\text{subject to} & \quad 0 \leq \lambda \leq \lambda_{\text{max}},
\end{align*}
\]

where
6 = \hat{b} - A_2 x_k \quad (10.1)

\hat{d} = A_2 d_k.

Summary of the Method of Zoutendijk (Case of Linear Constraints)

We summarize below Zoutendijk's method for minimizing a differentiable function $f$ in the presence of linear constraints of the form $Ax \leq b$ and $Qx = q$.

**Initialization Step** Find a starting feasible solution $x_1$ with $Ax_1 \leq b$ and $Qx_1 = q$. Let $k = 1$ and go to the Main Step.

**Main Step**

1. Given $x_k$, suppose that $A'$ and $b'$ are decomposed into $(A_1', A_2')$ and $(b_1', b_2')$ so that $A_1 x_k = b_1$ and $A_2 x_k < b_2$. Let $d_k$ be an optimal solution to the following problem (note that Problem P2 or P3 could be used instead):

\[
\begin{align*}
\text{Minimize} & \quad \nabla f(x)'d \\
\text{subject to} & \quad A_1 d \leq 0 \\
& \quad Qd = 0 \\
& \quad -1 \leq d_j \leq 1 \text{ for } j = 1, \ldots, n.
\end{align*}
\]

If $\nabla f(x_k)'d_k = 0$, stop; $x_k$ is KKT point, with the dual variables to the foregoing problem giving the corresponding Lagrange multipliers. Otherwise, go to Step 2.

2. Let $\lambda_k$ be an optimal solution to the following line search problem:

\[
\begin{align*}
\text{Minimize} & \quad f(x_k + \lambda d_k) \\
\text{subject to} & \quad 0 \leq \lambda \leq \lambda_{max},
\end{align*}
\]

where $\lambda_{max}$ is determined according to (10.1). Let $x_{k+1} = x_k + \lambda_k d_k$, identify the new set of binding constraints at $x_{k+1}$, and update $A_1$ and $A_2$ accordingly. Replace $k$ by $k + 1$ and go to Step 1.

**10.1.5 Example**

Consider the following problem:
Minimize $2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2$
subject to $x_1 + x_2 \leq 2$
$d_1$
subject to $d_2 \leq 0$
$-1 \leq d_1 \leq 1$
$-1 \leq d_2 \leq 1.$

Note that $\nabla f(x) = (4x_1 - 2x_2 - 4, 4x_2 - 2x_1 - 6)^t$. We solve the problem using Zoutendijk's procedure starting from the initial point $x_1 = (0, 0)^t$. Each iteration of the algorithm consists of the solution of a subproblem given in Step 1 to find the search direction and then a line search along this direction.

**Iteration 1:**

**Search Direction** At $x_1 = (0, 0)^t$ we have $\nabla f(x_1) = (-4, -6)^t$. Furthermore, at the point $x_1$, only the nonnegativity constraints are binding so that the index set of active constraints is given by $I = \{3, 4\}$. The direction-finding problem is given by:

Minimize $-4d_1 - 6d_2$
subject to $-d_1 \leq 0$
$d_1$
subject to $-d_2 \leq 0$
$-1 \leq d_1 \leq 1$
$-1 \leq d_2 \leq 1.$

This problem can be solved, for example, by the simplex method for linear programming; the optimal solution is $d_1 = (1, 1)^t$, and the optimal objective value for the direction-finding problem is $-10$. Figure 10.2 gives the feasible region for the subproblem, and the reader can readily verify geometrically that $(1, 1)$ is indeed the optimal solution.

![Figure 10.2 Iteration 1.](image-url)
Methods of Feasible Directions

Line Search  We now need to find a feasible point along the direction $(1, 1)$ starting from the point $(0, 0)$ having a minimum value of $f(x) = 2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2$. Any point along this direction can be written as $x_1 + \lambda d_1 = (\lambda, \lambda)\hat{f}$ and the objective function $f(x_1 + \lambda d_1) = -10\lambda + 2\lambda^2$. The maximum value of $\lambda$ for which $x_1 + \lambda d_1$ is feasible is computed using (10.1) and is given by

$$\lambda_{\text{max}} = \min\{2/2, 5/6\} = 5/6.$$

Hence, if $x_1 + \lambda_1 d_1$ is the new point, the value of $\lambda_1$ is obtained by solving the following one-dimensional search problem:

Minimize $-10\lambda + 2\lambda^2$
subject to $0 \leq \lambda \leq \frac{5}{6}$.

Since the objective function is convex and the unconstrained minimum occurs at $5/2$, the solution is $\lambda_1 = \frac{5}{6}$, so that $x_2 = x_1 + \lambda_1 d_1 = (5/6, 5/6)\hat{f}$.

Iteration 2:

Search Direction  At the point $x_2 = (5/6, 5/6)\hat{f}$, we have $\nabla f(x_2) = (-7/3, -13/3)\hat{f}$. Furthermore, the set of binding constraints at the point $x_2$ is given by $l = \{2\}$, so that the direction to move is obtained by solving the following problem:

Minimize $\frac{7}{3}d_1 - \frac{13}{3}d_2$
subject to $d_1 + 5d_2 \leq 0$
$-1 \leq d_1 \leq 1$
$-1 \leq d_2 \leq 1$.

The reader can verify from Figure 10.3 that the optimal solution to the above linear program is $d_2 = (1, -1/5)\hat{f}$, and the corresponding objective function value is $-22/15$.

Line Search  Starting from the point $x_2$, any point in the direction $d_2$ can be written as $x_2 + \lambda d_2 = (5/6 + \lambda, 5/6 - 1/5\lambda)$, and the corresponding objective function value is $f(x_2 + \lambda d_2) = -125/8 - 22/5\lambda + 62/25\lambda^2$. The maximum value of $\lambda$ for which $x_2 + \lambda d_2$ is feasible is obtained from (10.1) as
Therefore, $\lambda_2$ is the optimal solution to the following problem:

\[
\begin{align*}
\text{Minimize} & \quad \frac{125}{8} - \frac{22}{15} \lambda + \frac{62}{25} \lambda^2 \\
\text{subject to} & \quad 0 \leq \lambda \leq \frac{5}{12}.
\end{align*}
\]

The optimal solution is $\lambda_2 = 55/186$, the unconstrained minimizer of the objective function, so that $x_3 = x_2 + \lambda_2 d_2 = (35/31, 24/31)^t$.

**Iteration 3:**

**Search Direction** At $x_3 = (35/31, 24/31)^t$, we have $\nabla f(x_3) = (-32/31, -160/31)^t$. Furthermore, the set of binding constraints at the point $x_3$ is given $I = \{2\}$, so that the direction to move is obtained by solving the following problem:

\[
\begin{align*}
\text{Minimize} & \quad \frac{32}{31} d_1 - \frac{160}{31} d_2 \\
\text{subject to} & \quad d_1 + 5d_2 \leq 0 \\
& \quad -1 \leq d_1 \leq 1 \\
& \quad -1 \leq d_2 \leq 1.
\end{align*}
\]

\[
\lambda_{\text{max}} = \min \left\{ \frac{1}{3}, \frac{5}{6} \right\} = \frac{5}{12}.
\]

Figure 10.3 Iteration 2.
Methods of Feasible Directions

Figure 10.4 Termination at Iteration 3.

The reader can easily verify from Figure 10.4 that \( d_3 = (1, -1/5)^T \) indeed solves the above linear program, with the Lagrange multiplier associated with the first constraint being 32/31 and zero for the other constraints. The corresponding objective function value is zero, and the procedure is terminated. Furthermore, \( \bar{x} = x_3 = (35/31, 24/31)^T \) is a KKT point with the only nonzero Lagrange multiplier being associated with \( x_1 + 5x_2 \leq 5 \) and equal to 32/31. (Verify this graphically from Figure 10.5.) In this particular problem, \( f \) is convex, and by Theorem 4.3.8, \( \bar{x} \) is indeed the optimal solution.

Table 10.1 summarizes the computations for solving the problem. The progress of the algorithm is shown in Figure 10.5.

Problems Having Nonlinear Inequality Constraints

We now consider the following problem, where the feasible region is defined by a system of inequality constraints that are not necessarily linear:

\[
\begin{align*}
0 \leq x_1 \\
10 \leq x_2 \\
5 \leq x_1 + 5x_2 \\
5 \leq x_1 + 2x_2 \\
10 \leq 2x_1 + x_2 \\
20 \leq 5x_1 + 3x_2 \\
20 \leq x_1 + 10x_2 \\
20 \leq 5x_1 + x_2
\end{align*}
\]

Figure 10.5 Method of Zoutendijk for the case of linear constraints.
Table 10.1 Summary of Computations for the Method of Zoutendijk

<table>
<thead>
<tr>
<th>Iter. $k$</th>
<th>$x_k$</th>
<th>$f(x_k)$</th>
<th>Search Direction</th>
<th>Line Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$(0, 0)$</td>
<td>0</td>
<td>$(-4, -6)$</td>
<td>$(3, 4)$</td>
</tr>
<tr>
<td>2</td>
<td>$(\frac{5}{6}, \frac{5}{6})$</td>
<td>$-6.94$</td>
<td>$(\frac{7}{3}, -\frac{13}{3})$</td>
<td>$(2)$</td>
</tr>
<tr>
<td>3</td>
<td>$(\frac{35}{31}, \frac{24}{31})$</td>
<td>$-7.16$</td>
<td>$(\frac{32}{31}, -\frac{160}{31})$</td>
<td>$(2)$</td>
</tr>
</tbody>
</table>

Minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, ..., m$.

Theorem 10.1.6 gives a sufficient condition for a vector $d$ to be an improving feasible direction.

10.1.6 Theorem

Consider the problem to minimize $f(x)$ subject to $g_i(x) \leq 0$ for $i = 1, ..., m$. Let $x$ be a feasible solution, and let $I$ be the set of binding or active constraints, that is, $I = \{i : g_i(x) = 0\}$. Furthermore, suppose that $f$ and $g_i$ for $i \in I$ are differentiable at $x$ and that each $g_i$ for $i \notin I$ is continuous at $x$. If $\nabla f(x)'d < 0$ and $\nabla g_i(x)'d < 0$ for $i \in I$, then $d$ is an improving feasible direction.

Proof

Let $d$ satisfy $\nabla f(x)'d < 0$ and $\nabla g_i(x)'d < 0$ for $i \in I$. For $i \notin I$, $g_i(x) < 0$ and $g_i$ is continuous at $x$ so that $g_i(x + \lambda d) \leq 0$ for $\lambda > 0$ and small enough. By differentiability of $g_i$ for $i \in I$,

$$g_i(x + \lambda d) = g_i(x) + \lambda \nabla g_i(x)'d + \lambda \|d\| \alpha(x; \lambda d),$$

where $\alpha(x; \lambda d) \to 0$ as $\lambda \to 0$. Since $\nabla g_i(x)'d < 0$, then $g_i(x + \lambda d) < g_i(x) = 0$ for $\lambda > 0$ and small enough. Hence, $g_i(x + \lambda d) \leq 0$ for $i = 1, ..., m$; that is, $x + \lambda d$ is feasible, for $\lambda > 0$ and small enough. By a similar argument, since $\nabla f(x)'d < 0$, we get $f(x + \lambda d) < f(x)$ for $\lambda > 0$ and small enough. Hence, $d$ is an improving and feasible direction. This completes the proof.
Figure 10.6 illustrates the collection of improving feasible directions at $\bar{x}$. A vector $d$ satisfying $\nabla g_i(\bar{x})'d = 0$ is tangential to the set $\{x : g_i(x) = 0\}$ at $\bar{x}$. Because of the nonlinearity of $g_i$, moving along such a vector $d$ may lead to infeasible points, thus necessitating the strict inequality $\nabla g_i(\bar{x})'d < 0$.

To find a vector $d$ satisfying $\nabla f(x)'d < 0$ and $\nabla g_i(x)'d < 0$ for $i \in I$, it is only natural to minimize the maximum of $\nabla f(x)'d$ and $\nabla g_i(x)'d$ for $i \in I$. Denoting this maximum by $z$, and introducing the normalization restrictions $-1 \leq d_j \leq 1$ for each $j$, we get the following direction-finding problem:

Minimize $z$

subject to $\nabla f(x)'d - z \leq 0$

$\nabla g_i(x)'d - z \leq 0$ for $i \in I$

$-1 \leq d_j \leq 1$ for $j = 1, \ldots, n$.

Let $(\bar{z}, \bar{d})$ be an optimal solution to the above linear program. If $\bar{z} < 0$, then $\bar{d}$ is obviously an improving feasible direction. If, on the other hand, $\bar{z} = 0$, then the current vector is a Fritz John point, as demonstrated below.

---

Figure 10.6 Improving feasible directions for nonlinear constraints.
10.1.7 Theorem

Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \). Let \( x \) be a feasible solution, and let \( I = \{i : g_i(x) = 0\} \). Consider the following direction-finding problem:

\[
\begin{align*}
\text{Minimize} & \quad z \\
\text{subject to} & \quad \nabla f(x)'d - z \leq 0 \\
& \quad \nabla g_i(x)'d - z \leq 0 \quad \text{for } i \in I \\
& \quad -1 \leq d_j \leq 1 \quad \text{for } j = 1, \ldots, n.
\end{align*}
\]

Then, \( x \) is a Fritz John point if and only if the optimal objective value to the above problem is equal to zero.

**Proof**

The optimal objective value to the above problem is equal to zero if and only if the system \( \nabla f(x)'d < 0 \) and \( \nabla g_i(x)'d < 0 \) for \( i \in I \) has no solution. By Theorem 2.4.9, this system has no solution if and only if there exist scalars \( u_0 \) and \( u_i \) for \( i \in I \) such that

\[
\begin{align*}
u_0 \nabla f(x) + \sum_{i \in I} u_i \nabla g_i(x) &= 0 \\
u_0 &\geq 0, \quad u_i \geq 0 \quad \text{for } i \in I
\end{align*}
\]

either \( u_0 > 0 \) or else \( u_i > 0 \) for some \( i \in I \).

These are precisely the Fritz John conditions, and the proof is complete.

**Summary of the Method of Zoutendijk (Case of Nonlinear Inequality Constraints)**

*Initialization Step* Choose a starting point \( x_1 \) such that \( g_i(x_1) \leq 0 \) for \( i = 1, \ldots, m \). Let \( k = 1 \) and go to the Main Step.

*Main Step*

1. Let \( I = \{i : g_i(x_k) = 0\} \) and solve the following problem:

\[
\begin{align*}
\text{Minimize} & \quad z \\
\text{subject to} & \quad \nabla f(x_k)'d - z \leq 0 \\
& \quad \nabla g_i(x_k)'d - z \leq 0 \quad \text{for } i \in I \\
& \quad -1 \leq d_j \leq 1 \quad \text{for } j = 1, \ldots, n.
\end{align*}
\]
Let \((z_k, d_k)\) be an optimal solution. If \(z_k = 0\), stop; \(x_k\) is a Fritz John point. If \(z_k < 0\), go to Step 2.

2. Let \(\lambda_k\) be an optimal solution to the following line search problem:

\[
\begin{align*}
\text{Minimize} & \quad f(x_k + \lambda d_k) \\
\text{subject to} & \quad 0 \leq \lambda \leq \lambda_{\max},
\end{align*}
\]

where \(\lambda_{\max} = \sup\{\lambda : g_i(x_k + \lambda d_k) \leq 0 \text{ for } i = 1, \ldots, m\}\). Let \(x_{k+1} = x_k + \lambda_k d_k\), replace \(k\) by \(k + 1\), and go to Step 1.

10.1.8 Example

Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad 2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2 \\
\text{subject to} & \quad x_1 + 5x_2 \leq 5 \\
& \quad 2x_1^2 - x_2 \leq 0 \\
& \quad -x_1 - x_2 \leq 0.
\end{align*}
\]

We shall solve the problem using the method of Zoutendijk. The procedure is initiated from a feasible point \(x_1 = (0.00, 0.75)^t\). The reader may note that \(\nabla f(x) = (4x_1 - 2x_2 - 4, 4x_2 - 2x_1 - 6)^t\).

**Iteration 1:**

**Search Direction** At the point \(x_1 = (0.00, 0.75)^t\), we have \(\nabla f(x_1) = (-5.50, -3.00)^t\), and the binding constraints are defined by \(I = \{3\}\). We have \(\nabla g_3(x_1) = (-1, 0)^t\). The direction-finding problem is then given as follows:

\[
\begin{align*}
\text{Minimize} & \quad z \\
\text{subject to} & \quad -5.5d_1 - 3.0d_2 - z \leq 0 \\
& \quad -d_1 - z \leq 0 \\
& \quad -1 \leq d_j \leq 1 \quad \text{for } j = 1, 2.
\end{align*}
\]

Using the simplex method, for example, it can be verified that the optimal solution is \(d_1 = (1.00, -1.00)^t\) and \(z_1 = -1.00\).

**Line Search** Any point starting from \(x_1 = (0.00, 0.75)^t\) along the direction \(d_1 = (1.00, -1.00)^t\) can be written as \(x_1 + \lambda d_1 = (\lambda, 0.75 - \lambda)^t\), and the
corresponding value of the objective function is given by \( f(x_1 + \lambda d_1) = 6 \lambda^2 - 2.5 \lambda - 3.375 \). The reader can verify that the maximum value of \( \lambda \) for which \( x_1 + \lambda d_1 \) is feasible is given by \( \lambda_{\text{max}} = 0.4114 \), whence the constraint \( 2x_1^2 - x_2 \leq 0 \) becomes binding. The value of \( \lambda_1 \) is obtained by solving the following one-dimensional search problem:

\[
\begin{align*}
\text{Minimize} & \quad 6 \lambda^2 - 2.5 \lambda - 3.375 \\
\text{subject to} & \quad 0 \leq \lambda \leq 0.4114.
\end{align*}
\]

The optimal value can readily be found to be \( \lambda_1 = 0.2083 \). Hence, \( x_2 = (x_1 + \lambda_1 d_1) = (0.2083, 0.5417)' \).

**Iteration 2:**

**Search Direction** At the point \( x_2 = (0.2083, 0.5417)' \), we have \( \nabla f(x_2) = (-4.2500, -4.2500)' \). There are no binding constraints, and hence the direction-finding problem is given by:

\[
\begin{align*}
\text{Minimize} & \quad z \\
\text{subject to} & \quad -4.25d_1 - 4.25d_2 - z \leq 0 \\
& \quad -1 \leq d_j \leq 1 \quad \text{for} \ j = 1, 2.
\end{align*}
\]

The optimal solution is \( d_2 = (1, 1)' \) and \( z_2 = -8.50 \).

**Line Search** The reader can verify that the maximum value of \( \lambda \) for which \( x_2 + \lambda d_2 \) is feasible is \( \lambda_{\text{max}} = 0.3472 \), whence the constraint \( x_1 + 5x_2 \leq 5 \) becomes binding. The value of \( \lambda_2 \) is obtained by minimizing \( f(x_2 + \lambda d_2) = 2\lambda^2 - 8.5\lambda - 3.6354 \) subject to \( 0 \leq \lambda \leq 0.3472 \). This yields \( \lambda_2 = 0.3472 \), so that \( x_3 = x_2 + \lambda_2 d_2 = (0.5555, 0.8889)' \).

**Iteration 3:**

**Search Direction** At the point \( x_3 = (0.5555, 0.8889)' \), we have \( \nabla f(x_3) = (-3.5558, -3.5554)' \), and the binding constraints are defined by \( I = \{1\} \). The direction-finding problem is given by:

\[
\begin{align*}
\text{Minimize} & \quad z \\
\text{subject to} & \quad -3.5558d_1 - 3.5554d_2 - z \leq 0 \\
& \quad d_1 + 5d_2 - z \leq 0 \\
& \quad -1 \leq d_j \leq 1 \quad \text{for} \ j = 1, 2.
\end{align*}
\]
The optimal solution is $d_3 = (1.0000, -0.5325)'$ and $z_3 = -1.663$.

**Line Search** The reader can verify that the maximum value of $\lambda$ for which $x_3 + \lambda d_3$ is feasible is $\lambda_{\text{max}} = 0.09245$, whence the constraint $2x_1^2 - x_2 \leq 0$ becomes binding. The value of $\lambda_3$ is obtained by minimizing $f(x_3 + \lambda d_3) = 1.5021\lambda^2 - 5.4490\lambda - 6.3455$ subject to $0 \leq \lambda \leq 0.09245$. The optimal solution is $\lambda_3 = 0.09245$, so that $x_4 = x_3 + \lambda_3 d_3 = (0.6479, 0.8397)'$.

**Iteration 4:**

**Search Direction** At the point $x_4 = (0.6479, 0.8397)'$, we have $\nabla f(x_4) = (-3.0878, -3.9370)'$, and the binding constraints are defined by $I = \{2\}$. The direction-finding problem is as follows:

\[
\begin{align*}
\text{Minimize} & \quad z \\
\text{subject to} & \quad -3.0878d_1 - 3.9370d_2 - z \leq 0 \\
& \quad 2.5916d_1 - d_2 - z \leq 0 \\
& \quad -1 \leq d_j \leq 1 \quad \text{for } j = 1, 2.
\end{align*}
\]

The optimal solution is $d_4 = (-0.5171, 1.0000)'$ and $z_4 = -2.340$.

**Line Search** The reader can verify that the maximum value of $\lambda$ for which $x_4 + \lambda d_4$ is feasible is $\lambda_{\text{max}} = 0.0343$, whence the constraint $x_1 + 5x_2 \leq 5$ becomes binding. The value of $\lambda_4$ is obtained by minimizing $f(x_4 + \lambda d_4) = 3.569\lambda^2 - 2.340\lambda - 6.481$ subject to $0 \leq \lambda \leq 0.0343$, which gives $\lambda_4 = 0.0343$. Hence, the new point is $x_5 = x_4 + \lambda_4 d_4 = (0.6302, 0.8740)'$. The value of the objective function here is $-6.5443$, compared with the value of $-6.5590$ at the optimal point $(0.658872, 0.868226)'$.

Table 10.2 summarizes the computations for the first four iterations. Figure 10.7 depicts the progress of the algorithm. Note the zigzagging tendency of the algorithm, as might be expected because of the first-order approximations used by this method.

**Treatment of Nonlinear Equality Constraints**

The foregoing method of feasible directions must be modified to handle nonlinear equality constraints. To illustrate, consider Figure 10.8 for the case of a single equality constraint. Given a feasible point $x_k$, there exists no nonzero direction $d$ such that $h(x_k + \lambda d) = 0$ for $\lambda \in (0, \delta)$, for some positive $\delta$. This
### Table 10.2 Summary of Computations for the Method of Zoutendijk for the Case of Nonlinear Constraints

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$x_k$</th>
<th>$f(x_k)$</th>
<th>$\nabla f(x_k)$</th>
<th>$d_k$</th>
<th>$z_k$</th>
<th>$\lambda_{max}$</th>
<th>$\lambda_i$</th>
<th>$x_{k+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.00, 0.75)</td>
<td>-3.3750</td>
<td>(-5.50, -3.00)</td>
<td>(1.0000, -1.0000)</td>
<td>-1.000</td>
<td>0.4140</td>
<td>0.2083</td>
<td>(0.2083, 0.5417)</td>
</tr>
<tr>
<td>2</td>
<td>(0.2083, 0.5477)</td>
<td>-3.6354</td>
<td>(-4.25, -4.25)</td>
<td>(1.0000, 1.0000)</td>
<td>-8.500</td>
<td>0.3472</td>
<td>0.3472</td>
<td>(0.55555, 0.8889)</td>
</tr>
<tr>
<td>3</td>
<td>(0.5555, 0.8889)</td>
<td>-6.3455</td>
<td>(-3.5558, -3.5554)</td>
<td>(1.0000, -0.5325)</td>
<td>-1.663</td>
<td>0.09245</td>
<td>0.09245</td>
<td>(0.6479, 0.8397)</td>
</tr>
<tr>
<td>4</td>
<td>(0.6479, 0.8397)</td>
<td>-6.4681</td>
<td>(-3.0878, -3.9370)</td>
<td>(-0.5171, 1.0000)</td>
<td>-2.340</td>
<td>0.0343</td>
<td>0.0343</td>
<td>(0.6302, 0.8740)</td>
</tr>
</tbody>
</table>
difficulty may be overcome by moving along a tangential direction $d_k$ having
\[ \nabla h(x_k) \cdot d_k = 0, \]
and then making a corrective move back to the feasible region.

To be more specific, consider the following problem:

Minimize \[ f(x) \]
subject to \[ g_i(x) \leq 0 \quad \text{for} \quad i = 1, \ldots, m \]
\[ h_i(x) = 0 \quad \text{for} \quad i = 1, \ldots, \ell. \]

Let $x_k$ be a feasible point, and let $I = \{i : g_i(x_k) = 0\}$. Solve the following linear program:

\[ \nabla h(x_k) \]

Figure 10.8 Nonlinear equality constraints.
Minimize $\nabla f(x_k)^T d$

subject to $\nabla g_i(x_k)^T d \leq 0$ for $i \in I$

$\nabla h_i(x_k)^T d = 0$ for $i = 1, \ldots, \ell$.

The resulting direction $d_k$ is tangential to the equality constraints and to some of the binding nonlinear inequality constraints. A search along $d_k$ is used, and then a move back to the feasible region leads to $x_{k+1}$ and the process is repeated.

**Use of Near-Binding Constraints**

Recall that the direction-finding problem, for both linear and nonlinear inequality-constrained problems, used only the set of binding constraints. If a given point is close to the boundary of one of the constraints, and if this constraint is not used in the process of finding a direction of movement, it is possible that only a small step could be taken before we hit the boundary of this constraint. In Figure 10.9, the only binding restriction at $x$ is the first constraint. However, $x$ is close to the boundary of the second constraint. If the set $I$ in the direction-finding Problem $P$ is taken as $I = \{1\}$, then the optimal direction will be $d$, and only a small movement can be realized before the boundary of Constraint 2 is reached. If, on the other hand, both Constraints 1 and 2 are treated as being active so that $I = \{1, 2\}$, then the direction-finding Problem $P$ will produce the direction $d$, thus providing more room to move before reaching the boundary of the feasible region. Therefore, it is suggested to let the set $I$ be the collection of near-binding constraints. More precisely, $I$ is taken as $\{i : g_i(x) + \varepsilon \geq 0\}$ rather than $\{i : g_i(x) = 0\}$, where $\varepsilon > 0$ is a suitable small scalar. Of course,
some caution is required in such a construct to prevent premature termination. As we discuss in detail in Section 10.2, the method of feasible directions presented in this section does not necessarily converge to a Fritz John point. This results from the fact that the algorithmic map is not closed. Through a more formal use of the concept of near-binding constraints presented here, closedness of the algorithmic map, and hence convergence of the overall algorithm, can be established.

10.2 Convergence Analysis of the Method of Zoutendijk

In this section we discuss the convergence properties of Zoutendijk's method of feasible directions presented in Section 10.1. As we shall learn shortly, the algorithmic map of Zoutendijk's method is not closed, and hence convergence is not generally guaranteed. A modification of the method credited to Topkis and Veinott [1967] assures convergence of the algorithm to a Fritz John point.

Note that the algorithmic map $A$ of the method of Zoutendijk is composed of the maps $M$ and $D$. The direction-finding map $D: \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ is defined by $(x, d) \in D(x)$ if $d$ is an optimal solution to one of the direction-finding Problems P1, P2, or P3 discussed in Section 10.1. The line search map $M: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is defined by $y \in M(x, d)$ if $y$ is an optimal solution to the problem to minimize $f(x + \lambda d)$ subject to $\lambda \geq 0$ and $x + \lambda d \in S$, where $S$ is the feasible region. We demonstrate below that the map $D$ is not closed in general.

10.2.1 Example (D Is Not Closed)

Consider the following problem:

Minimize $-2x_1 - x_2$
subject to $x_1 + x_2 \leq 2$
$x_1, x_2 \geq 0$.

The problem is illustrated in Figure 10.10. Consider the sequence of vectors $\{x_k\}$, where $x_k = (0, 2-1/k)'$. Note that at each $x_k$, the only binding constraint is $x_1 \geq 0$, and the direction-finding problem is given by:

Minimize $-2d_1 - d_2$
subject to $0 \leq d_1 \leq 1$
$-1 \leq d_2 \leq 1$.

The optimal solution $d_k$ to the above problem is obviously $(1, 1)'$. At the limit point $x = (0, 2)'$, however, the restrictions $x_1 \geq 0$ and $x_1 + x_2 \leq 2$ are both binding, so that the direction-finding problem is given by:
Figure 10.10 The direction-finding map $D$ is not closed.

Minimize $-2d_1 - d_2$
subject to $d_1 + d_2 \leq 0$
$0 \leq d_1 \leq 1$
$-1 \leq d_2 \leq 1$.

The optimal solution $\tilde{d}$ to the above problem is given by $(1, -1)'$. Thus,

$x_k \rightarrow x$

$(x_k, d_k) \rightarrow (x, d)$,

where $d = (1, 1)'$. Since $D(x) = \{(x, \tilde{d})\}$, $(x, d) \notin D(x)$. Therefore, the direction-finding map $D$ is not closed at $x$.

The line search map $M: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is used by all feasible direction algorithms for solving a problem of the form to minimize $f(x)$ subject to $x \in S$. Given a feasible point $x$ and an improving feasible direction $d$, $y \in M(x, d)$ means that $y$ is an optimal solution to the problem to minimize $f(x + \lambda d)$ subject to $\lambda \geq 0$ and $x + \lambda d \in S$. In Example 10.2.2 we demonstrate that this map is not closed. The difficulty here is that the possible step length that could be taken before leaving the feasible region may approach zero, causing jamming.

10.2.2 Example (M Is Not Closed)

Consider the following problem:

Minimize $2x_1 - x_2$
subject to $(x_1, x_2) \in S$,
where \( S = \{(x_1, x_2) : x_1^2 + x_2^2 \leq 1\} \cup \{(x_1, x_2) : |x_1| \leq 1, 0 \leq x_2 \leq 1\} \). The problem is illustrated in Figure 10.11, and the optimal point \( \bar{x} \) is given by \((-1, 1)^t\). Now consider the sequence \( \{(x_k, d_k)\} \) formed as follows. Let \( x_1 = (1, 0)^t \) and \( d_1 = (-1/\sqrt{2}, -1/\sqrt{2})^t \). Given \( x_k \), the next iterate \( x_{k+1} \) is given by moving along \( d_k \) until the boundary of \( S \) is reached. Given \( x_{k+1} \), the next direction \( d_{k+1} \) is taken as \( (\xi - x_{k+1})/\|\xi - x_{k+1}\| \), where \( \xi \) is the point on the boundary of \( S \) that is equidistant from \( x_{k+1} \) and \((-1, 0)^t\).

The sequence \( \{(x_k, d_k)\} \) is shown in Figure 10.11 and obviously converges to \((x, d)\), where \( x = (-1, 0)^t \) and \( d = (0, 1)^t \). The line search map \( M \) is defined by \( y_k \in M(x_k, d_k) \) if \( y_k \) is an optimal solution to the problem to minimize \( f(x_k + \lambda d_k) \) subject to \( \lambda \geq 0 \) and \( x_k + \lambda d_k \in S \). Obviously, \( y_k = x_{k+1} \) and, hence, \( y_k \to x \). Thus,

\[
(x_k, d_k) \to (x, d) \quad y_k \to x \quad \text{where} \quad y_k \in M(x_k, d_k).
\]

However, minimizing \( f \) starting from \( x \) in the direction \( d \) yields \( \bar{x} \), so that \( x \not\in M(x, d) \). Thus, \( M \) is not closed at \((x, d)\).

---

**Figure 10.11** The line search map \( M \) is not closed.
Wolfe's Counterexample

We demonstrated above that both the direction-finding map and the line search map of Zoutendijk are not closed. Example 10.2.3 shows that Zoutendijk's algorithm may not converge to a KKT point. The difficulty here is that the distance moved along the generated directions tends toward zero, causing jamming at a nonoptimal point.

10.2.3 Example (Wolfe [1972])

Consider the following problem:

Minimize \( \frac{4}{3}(x_1^2 - x_1 x_2 + x_2^2)^{3/4} - x_3 \)
subject to \(-x_1, -x_2, -x_3 \leq 0\)
\( x_3 \leq 2. \)

Note that the objective function is convex and that the optimal solution is achieved at the unique point \( \bar{x} = (0, 0, 2)' \). We solve this problem using Zoutendijk's procedure, starting from the feasible point \( x_1 = (0, a, 0)' \), where \( a \leq 1/(2\sqrt{2}) \). Given a feasible point \( x_k \), the direction of movement \( d_k \) is obtained by solving the following Problem P2:

Minimize \( \nabla f(x_k)'d \)
subject to \( A_1d \leq 0 \)
\( d'd \leq 1, \)

where \( A_1 \) is the matrix whose rows are the gradients of the binding constraints at \( x_k \). Here, \( x_1 = (0, a, 0)' \) and \( \nabla f(x_1) = (-\sqrt{a}, 2\sqrt{a}, -1)' \). Note that

\[
A_1 = \begin{bmatrix}
-1 & 0 & 0 \\
0 & 0 & -1
\end{bmatrix}
\]

and that the optimal solution to Problem P2 above is \( d_1 = -\nabla f(x_1)/\|\nabla f(x_1)\| \).

The optimal solution \( \lambda_1 \) to the line search problem to minimize \( f(x_1 + \lambda d_1) \)
subject to \( \lambda \geq 0 \) and \( x_1 + \lambda d_1 \in S \) yields \( x_2 = x_1 + \lambda_1 d_1 = ((1/2)a, 0, (1/2)\sqrt{a}) \).

Repeating this process, we obtain the sequence \( \{x_k\} \), where
Methods of Feasible Directions

Note that this sequence converges to the point \( \hat{x} = [0, 0, (1 + (1/2)^k) \sqrt{a}]^T \).
Since the optimal solution \( \bar{x} \) is unique, Zoutendijk's method converges to a point \( \hat{x} \) that is neither optimal nor a KKT point.

Topkis-Veinott’s Modification of the Feasible Direction Algorithm

We now describe a modification of Zoutendijk’s method of feasible directions. This modification was proposed by Topkis and Veinott [1967] and guarantees convergence to a Fritz John point. The problem under consideration is given by:

\[
\text{Minimize } f(x) \\
\text{subject to } g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m.
\]

Generating a Feasible Direction

Given a feasible point \( x \), a direction \( d \) is found by solving the following direction-finding linear programming problem \( \text{DF}(x) \):

\[
\text{Problem } \text{DF}(x): \text{Minimize } z \\
\text{subject to } \nabla f(x)^T d - z \leq 0 \\
\nabla g_i(x)^T d - z \leq -g_i(x) \quad \text{for } i = 1, \ldots, m \\
-1 \leq d_j \leq 1 \quad \text{for } j = 1, \ldots, n.
\]

Here, both binding and nonbinding constraints play a role in determining the direction of movement. As opposed to the method of feasible directions of Section 10.1, no sudden change in the direction is encountered when approaching the boundary of a currently nonbinding constraint.

Summary of the Method of Feasible Directions of Topkis and Veinott

A summary of the method of Topkis and Veinott for solving the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \) is given below. As will be shown later, the method converges to a Fritz John point.

Initialization Step  Choose a point \( x_1 \) such that \( g_i(x_1) \leq 0 \) for \( i = 1, \ldots, m \). Let \( k = 1 \) and go to the Main Step.
Main Step

1. Let \((z_k, d_k)\) be an optimal solution to the following linear programming problem:

Minimize \(z\)

subject to \(\nabla f(x_k)' d - z \leq 0\)

\(\nabla g_i(x_k)' d - z \leq -g_i(x_k)\) for \(i = 1, \ldots, m\)

\(-1 \leq d_j \leq 1\) for \(j = 1, \ldots, n\).

If \(z_k = 0\), stop; \(x_k\) is a Fritz John point. Otherwise, \(z_k < 0\), and go to Step 2.

2. Let \(\lambda_k\) be an optimal solution to the following line search problem:

Minimize \(f(x_k + \lambda d_k)\)

subject to \(0 \leq \lambda \leq \lambda_{\text{max}}\),

where \(\lambda_{\text{max}} = \sup \{ \lambda : g_i(x_k + \lambda d_k) \leq 0 \text{ for } i = 1, \ldots, m \}\). Let \(x_{k+1} = x_k + \lambda_k d_k\), replace \(k\) by \(k + 1\), and go to Step 1.

10.2.4 Example

Consider the following problem:

Minimize \(2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2\)

subject to \(x_1 + 5x_2 \leq 5\)

\(2x_1^2 - x_2 \leq 0\)

\(-x_1 \leq 0\)

\(-x_2 \leq 0\).

We go through five iterations of the algorithm of Topkis and Veinott, starting from the point \(x_1 = (0.00, 0.75)'\). Note that the gradient of the objective function is given by \(\nabla f(x) = (4x_1 - 2x_2 - 4, 4x_2 - 2x_1 - 6)'\), and the gradients of the constraint functions are \((1, 5)', (4x_1, -1)', (-1, 0)', and (0, -1)'\), which are used in defining the direction-finding problem at each iteration.

Iteration 1:

Search Direction At \(x_1 = (0.00, 0.75)'\), we have \(\nabla f(x_1) = (-5.5, -3.0)'\). Hence the direction-finding problem is as follows:
Minimize $z$
subject to $-5.5d_1 - 3d_2 - z \leq 0$
$\quad d_1 + 5d_2 - z \leq 1.25$
$\quad -d_2 - z \leq 0.75$
$\quad -d_1 - z \leq 0$
$\quad -d_2 - z \leq 0.75$
$\quad -1 \leq d_j \leq 1 \quad \text{for } j = 1, 2.$

The right-hand sides of the constraints 2 to 5 are $-g_i(x_1)$ for $i = 1, 2, 3, 4$. Note that one of the constraints, $-d_2 - z \leq 0.75$, is redundant. The optimal solution to the above problem is $d_1 = (0.7143, -0.03571)^t$ and $z_1 = 4.7143$.

**Line Search**  The reader can readily verify that the maximum value of $\lambda$ for which $x_1 + \lambda d_1$ is feasible is given by $\lambda_{\max} = 0.84$ and that $f(x_1 + \lambda d_1) = 0.972\lambda^2 - 4.036\lambda - 3.375$. Moreover, $\lambda_1 = 0.84$ solves the problem to minimize $f(x_1 + \lambda d_1)$ subject to $0 \leq \lambda \leq 0.84$. We then have $x_2 = x_1 + \lambda_1 d_1 = (0.60, 0.72)^t$.

**Iteration 2:**

**Search Direction**  At the point $x_2$ we have $\nabla f(x_2) = (-3.04, -4.32)^t$. The direction $d_2$ is obtained as the optimal solution to the problem:

Minimize $z$
subject to $-3.04d_1 - 4.32d_2 - z \leq 0$
$\quad d_1 + 5d_2 - z \leq 0.8$
$\quad 2.4d_1 - d_2 - z \leq 0$
$\quad -d_1 - z \leq 0.6$
$\quad -d_2 - z \leq 0.72$
$\quad -1 \leq d_j \leq 1 \quad \text{for } j = 1, 2.$

The optimal solution is $d_2 = (-0.07123, 0.1167)^t$ and $z_2 = -0.2877$.

**Line Search**  The maximum value of $\lambda$ such that $x_2 + \lambda d_2$ is feasible is given by $\lambda_{\max} = 1.561676$. The reader can easily verify that $f(x_2 + \lambda_2 d_2) = 0.054\lambda^2 - 0.2876\lambda - 5.8272$ attains a minimum over the interval $0 \leq \lambda \leq 1.561676$ at the point $\lambda_2 = 1.561676$. Hence, $x_3 = x_2 + \lambda_2 d_2 = (0.4888, 0.9022)^t$.

The process is then repeated. Table 10.3 summarizes the computations for five iterations. The progress of the algorithm is depicted in Figure 10.12. Note
### Table 10.3 Summary of the Topkis–Veinott Method

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$x_k$</th>
<th>$f(x_k)$</th>
<th>$\nabla f(x_k)$</th>
<th>$d_k$</th>
<th>$z_k$</th>
<th>$\lambda_{max}$</th>
<th>$\lambda_k$</th>
<th>$x_{k+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.0000, 0.7500)</td>
<td>-3.3750</td>
<td>(-5.50, -3.00)</td>
<td>(0.7143, -0.03571)</td>
<td>-0.7143</td>
<td>0.84</td>
<td>0.84</td>
<td>(0.6000, 0.7200)</td>
</tr>
<tr>
<td>2</td>
<td>(0.6000, 0.7200)</td>
<td>-5.8272</td>
<td>(-3.04, -4.32)</td>
<td>(-0.07123, 0.1167)</td>
<td>-0.2877</td>
<td>1.561676</td>
<td>1.561676</td>
<td>(0.4888, 0.9022)</td>
</tr>
<tr>
<td>3</td>
<td>(0.4888, 0.9022)</td>
<td>-6.1446</td>
<td>(-3.8492, -3.3688)</td>
<td>(0.09574, -0.05547)</td>
<td>-0.1816</td>
<td>1.56395</td>
<td>1.56395</td>
<td>(0.6385, 0.8154)</td>
</tr>
<tr>
<td>4</td>
<td>(0.6385, 0.8154)</td>
<td>-6.3425</td>
<td>(-5.6308, -4.0154)</td>
<td>(-0.01595, 0.04329)</td>
<td>-0.0840</td>
<td>1.41895</td>
<td>1.41895</td>
<td>(0.6159, 0.8768)</td>
</tr>
<tr>
<td>5</td>
<td>(0.6159, 0.8768)</td>
<td>-6.5082</td>
<td>(-3.2900, -3.7246)</td>
<td>(0.02676, -0.01316)</td>
<td>-0.0303</td>
<td>1.45539</td>
<td>1.45539</td>
<td>(0.6548, 0.8575)</td>
</tr>
</tbody>
</table>
that at the end of five iterations, the point \((0.6548, 0.8575)^T\) is reached, having an objective function value of \(-6.5590\). Note that the optimal point is \((0.658872, 0.868226)^T\), with objective function value \(-6.613086\). Again, observe the zigzagging of the iterates generated by the algorithm.

**Convergence of the Method of Topkis and Veinott**

Theorem 10.2.7 establishes the convergence of the method of Topkis and Veinott to a Fritz John point. Two intermediate results are needed. Theorem 10.2.5 provides a necessary and sufficient condition for arriving at a Fritz John point and shows that an optimal solution to the direction-finding problem indeed provides an improving feasible direction.

**10.2.5 Theorem**

Let \(x\) be a feasible solution to the problem to minimize \(f(x)\) subject to \(g_i(x) \leq 0\) for \(i = 1, \ldots, m\). Let \((\bar{z}, \bar{d})\) be an optimal solution to problem DF(x). If \(\bar{z} < 0\), then \(\bar{d}\) is an improving feasible direction. Also, \(\bar{z} = 0\) if and only if \(x\) is a Fritz John point.

![Figure 10.12 Method of Topkis and Veinott.](image-url)
Proof

Let \( I = \{ i : g_i(x) = 0 \} \), and suppose that \( \overline{z} < 0 \). Examining Problem \( \text{DF}(x) \), we note that \( \nabla g_i(x)^T \overline{d} < 0 \) for \( i \in I \). This, together with the fact that \( g_i(x) < 0 \) for \( i \notin I \), implies that \( x + \lambda \overline{d} \) is feasible for \( \lambda > 0 \) and sufficiently small. Thus, \( \overline{d} \) is a feasible direction. Furthermore, \( \nabla f(x)^T \overline{d} < 0 \), and hence \( \overline{d} \) is an improving direction.

Now, we prove the second part of the theorem. Noting that \( g_i(x) = 0 \) for \( i \in I \) and that \( g_i(x) < 0 \) for \( i \notin I \), it can easily be verified that \( \overline{z} = 0 \) if and only if the system \( \nabla f(x)^T d < 0 \) and \( \nabla g_i(x)^T d < 0 \) for \( i \in I \) has no solution. By Theorem 2.4.9 this system has no solution if and only if \( x \) is a Fritz John point, and the proof is complete.

Lemma 10.2.6 will be used to prove Theorem 10.2.7, which establishes the convergence of the algorithm of Topkis and Veinott. The lemma states essentially that any feasible direction algorithm cannot generate a sequence of points and directions satisfying properties 1 through 4 stated below.

10.2.6 Lemma

Let \( S \) be a nonempty closed set in \( \mathbb{R}^n \), and let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) be continuously differentiable. Consider the problem to minimize \( f(x) \) subject to \( x \in S \). Furthermore, consider any feasible direction algorithm whose map \( A = MD \) is defined as follows. Given \( x, (x, d) \in D(x) \) means that \( d \) is an improving feasible direction of \( f \) at \( x \). Furthermore, \( y \in M(x, d) \) means that \( y = x + \lambda d \), where \( \lambda \) solves the line search problem to minimize \( f(x + \lambda d) \) subject to \( \lambda \geq 0 \) and \( x + \lambda d \in S \). Let \( \{ x_k \} \) be any sequence generated by such an algorithm, and let \( \{ d_k \} \) be the corresponding sequence of directions. Then there cannot exist a subsequence \( \{ (x_k, d_k) \} \) satisfying all the following properties:

1. \( x_k \rightarrow x \) for \( k \in \mathcal{K} \)
2. \( d_k \rightarrow d \) for \( k \in \mathcal{K} \)
3. \( x_k + \lambda d_k \in S \) for all \( \lambda \in [0, \delta] \) and for each \( k \in \mathcal{K} \) for some \( \delta > 0 \).
4. \( \nabla f(x)^T d < 0 \).

Proof

Suppose, by contradiction, that there exists a subsequence \( \{ (x_k, d_k) \} \) satisfying conditions 1 through 4. By Condition 4, there exists an \( \varepsilon > 0 \) such that \( \nabla f(x)^T d = -2\varepsilon \). Since \( x_k \rightarrow x \) and \( d_k \rightarrow d \) for \( k \in \mathcal{K} \), and since \( f \) is continuously differentiable, there exists a \( \delta' > 0 \) such that
\[ \nabla f(x_k + \lambda d_k)' d_k \leq -\varepsilon \quad \text{for } \lambda \in [0, \delta'] \text{ and for } k \in \mathcal{H} \text{ sufficiently large.} \quad (10.2) \]

Now, let \( \tilde{\delta} = \min\{\delta', \delta\} > 0. \) Consider \( k \in \mathcal{H} \) sufficiently large. By Condition 3, and by the definition of \( x_{k+1} \), we must have \( f(x_{k+1}) \leq f(x_k + \tilde{\delta} d_k) \). By the mean value theorem, \( f(x_k + \tilde{\delta} d_k) = f(x_k) + \tilde{\delta} \nabla f(x_k)' d_k \), where \( \tilde{x}_k = x_k + \lambda_k \tilde{\delta} d_k \) and \( \lambda_k \in (0, 1) \). By (10.2) it then follows that
\[
f(x_{k+1}) < f(x_k) - \varepsilon \tilde{\delta} \quad \text{for } k \in \mathcal{H} \text{ sufficiently large.} \quad (10.3)\]

Since the feasible direction algorithm generates a sequence of points having decreasing objective values, \( \lim_{k \to \infty} f(x_k) = f(x) \). In particular, both \( f(x_{k+1}) \) and \( f(x_k) \) approach \( f(x) \) if \( k \in \mathcal{H} \) approaches \( \infty \). Thus, from (10.3), we get \( f(x) \leq f(x) - \varepsilon \tilde{\delta} \), which is impossible, since \( \varepsilon, \tilde{\delta} > 0 \). This contradiction shows that no subsequence satisfying properties 1 through 4 could exist.

### 10.2.7 Theorem

Let \( f, g_i : \mathbb{R}^n \to \mathbb{R} \) for \( i = 1, \ldots, m \) be continuously differentiable, and consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \). Suppose that the sequence \( \{x_k\} \) is generated by the algorithm of Topkis and Veinott. Then any accumulation point of \( \{x_k\} \) is a Fritz John point.

**Proof**

Let \( \{x_k\} \in \mathcal{H} \) be a convergent subsequence with limit \( x \). We need to show that \( x \) is a Fritz John point. Suppose, by contradiction, that \( x \) is not a Fritz John point, and let \( \bar{z} \) be the optimal objective value of Problem \( DF(x) \). By Theorem 10.2.5, there exists an \( \varepsilon > 0 \) such that \( \bar{z} = -2\varepsilon \). For \( k \in \mathcal{H} \), consider Problem \( DF(x_k) \) and let \( (z_k, d_k) \) be an optimal solution. Since \( \{d_k\} \in \mathcal{H}' \) is bounded, there exists a subsequence \( \{d_k\} \) with limit \( d \). Furthermore, since \( f \) and \( g_i \) for \( i = 1, \ldots, m \) are continuously differentiable, and since \( x_k \to x \) for \( k \in \mathcal{H}' \), it follows that \( z_k \to \bar{z} \). In particular, for \( k \in \mathcal{H}' \) sufficiently large, we must have \( z_k < -\varepsilon \).

By the definition of Problem \( DF(x_k) \), we must have
\[
\nabla f(x_k)' d_k \leq z_k < -\varepsilon \quad \text{for } k \in \mathcal{H}' \text{ sufficiently large} \quad (10.4)
\]

\[
g_i(x_k) + \nabla g_i(x_k)' d_k \leq z_k < -\varepsilon \quad \text{for } k \in \mathcal{H}' \text{ sufficiently large,}
\]

for \( i = 1, \ldots, m \). \quad (10.5)

By the continuous differentiability of \( f \), (10.4) implies that \( \nabla f(x)' d < 0 \).
Since $g_i$ is continuously differentiable, from (10.5) there exists a $\delta > 0$ such that the following inequality holds true for each $\lambda \in [0, \delta]$:

$$g_i(x_k) + \nabla g_i(x_k + \lambda d_k)^T d_k < -\frac{\varepsilon}{2}$$

for $k \in \mathcal{K}'$ sufficiently large, for $i = 1, \ldots, m$. (10.6)

Now, let $\lambda \in [0, \delta]$. By the mean value theorem, and since $g_i(x_k) \leq 0$ for each $k$ and each $i$, we get

$$g_i(x_k + \lambda d_k) = g_i(x_k) + \lambda \nabla g_i(x_k + \alpha_{ik} \lambda d_k)^T d_k$$

$$= (1 - \lambda) g_i(x_k) + \lambda [g_i(x_k) + \nabla g_i(x_k + \alpha_{ik} \lambda d_k)^T d_k]$$

$$\leq \lambda [g_i(x_k) + \nabla g_i(x_k + \alpha_{ik} \lambda d_k)^T d_k]$$

(10.7)

where $\alpha_{ik} \in (0, 1)$. Since $\alpha_{ik} \lambda \in [0, \delta]$, from (10.6) and (10.7), it follows that $g_i(x_k + \lambda d_k) \leq -\lambda \varepsilon / 2 \leq 0$ for $k \in \mathcal{K}'$ sufficiently large and for $i = 1, \ldots, m$. This shows that $x_k + \lambda d_k$ is feasible for each $\lambda \in [0, \delta]$, for all $k \in \mathcal{K}'$ sufficiently large.

To summarize, we have exhibited a sequence $\{(x_k, d_k)\}_{\mathcal{K}'}$ that satisfies conditions 1 through 4 of Lemma 10.2.6. By the lemma, however, the existence of such a sequence is not possible. This contradiction shows that $x$ is a Fritz John point, and the proof is complete.

### 10.3 Successive Linear Programming Approach

In our foregoing discussion of Zoutendijk’s algorithm and its convergent variant as proposed by Topkis and Veinott, we have learned that at each iteration of this method, we solve a direction-finding linear programming problem based on first-order functional approximations in a minimax framework, and then conduct a line search along this direction. Conceptually, this is similar to successive linear program (SLP) approaches, also known as sequential, or recursive, linear programming. Here, at each iteration $k$, a direction-finding linear program is formulated based on first-order Taylor series approximations to the objective and constraint functions, in addition to appropriate step bounds or trust region restrictions on the direction components. If $d_k = 0$ solves this problem, then the current iterate $x_k$ is optimal to the first-order approximation, so, from Theorem 4.2.15, this solution is a KKT point and we terminate the procedure. Otherwise, the procedure either accepts the new iterate $x_{k+1} = x_k + d_k$ or rejects this iterate and reduces the step bounds, and then repeats this process. The decision as to whether to accept or reject the new iterate is typically made based on a merit function fashioned around the $\ell_1$, or the absolute value, penalty function [see Equation (9.8)].
The philosophy of this approach was introduced by Griffith and Stewart of the Shell Development Company in 1961, and has been widely used since then, particularly in the oil and chemical industries (see Exercise 10.53). The principal advantage of this type of method is its ease and robustness in implementation for large-scale problems, given an efficient and stable linear programming solver. As can be expected, if the optimum is a vertex of the (linearized) feasible region, rapid convergence is obtained. Indeed, once the algorithm enters a relatively close neighborhood of such a solution, it essentially behaves like Newton's algorithm applied to the binding constraints (under suitable regularity assumptions), with the Newton iterate being the (unique) linear programming solution, and a quadratic convergence rate obtains. Hence, highly constrained nonlinear programming problems that have nearly as many linearly independent active constraints as variables are very suitable for this class of algorithms. Real-world nonlinear refinery models tend to be of this nature, and problems having up to 1000 rows have been solved successfully. On the negative side, SLP algorithms exhibit slow convergence to nonvertex solutions, and they also have the disadvantage of violating nonlinear constraints en route to optimality.

Below, we describe an SLP algorithm, called the penalty successive linear programming (PSLP) algorithm, which employs the $\ell_1$ penalty function more actively in the direction-finding problem itself, rather than as only a merit function, and enjoys good robustness and convergence properties. The problem we consider is of the form:

\[ P: \text{Minimize } f(x) \]
\[ \text{subject to } \begin{array}{l}
  g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m \\
  h_i(x) = 0 \quad \text{for } i = 1, \ldots, \ell \\
  x \in X = \{x : Ax \leq b\},
\end{array} \tag{10.8} \]

where all functions are assumed to be continuously differentiable, $x \in \mathbb{R}^n$, and where the linear constraints defining the problem have all been accommodated into the set $X$.

Now let $F_E(x)$ be the $\ell_1$, or absolute value, exact penalty function of Equation (9.8), restated below for a penalty parameter $\mu > 0$:

\[ F_E(x) = f(x) + \mu \left( \sum_{i=1}^{m} \max\{0, g_i(x)\} + \sum_{i=1}^{\ell} |h_i(x)| \right). \]

Accordingly, consider the following (linearly constrained) penalty problem PP:

\[ PP: \text{Minimize } \{F_E(x) : x \in X\}. \tag{10.9a} \]

Substituting $y_i$ for $\max\{0, g_i(x)\}$, $i = 1, \ldots, m$, and writing $h_i(x)$ as the difference $z_i^+ - z_i^-$ of two nonnegative variables, where $|h_i(x)| = z_i^+ + z_i^-$, for $i = 1, \ldots,
We can equivalently rewrite (10.9a) without the nondifferentiable terms as follows:

\[
\text{PP: Minimize } f(x) + \mu \left[ \sum_{i=1}^{m} y_i + \sum_{i=1}^{\ell} (z_i^+ + z_i^-) \right]
\]

subject to \( y_i \geq g_i(x) \) for \( i = 1, \ldots, m \)
\( z_i^+ - z_i^- = h_i(x) \) for \( i = 1, \ldots, \ell \)
\( x \in X, y_i \geq 0 \) for \( i = 1, \ldots, m \)
\( z_i^+ \text{ and } z_i^- \geq 0 \) for \( i = 1, \ldots, \ell \).

(10.9b)

Note that given any \( x \in X \), since \( \mu > 0 \), the optimal completion \((y, z^+, z^-) = (y_1, \ldots, y_m, z_1^+, \ldots, z_\ell^+, z_1^-, \ldots, z_\ell^-)\) is determined by letting

\[
y_i = \max \{0, g_i(x)\}, \quad i = 1, \ldots, m
\]
\[
z_i^+ = \max \{0, h_i(x)\}, \quad z_i^- = \max \{0, -h_i(x)\}, \quad i = 1, \ldots, \ell
\]

so that \((z_i^+ + z_i^-) = |h_i(x)|, \quad i = 1, \ldots, \ell\). (10.10)

Consequently, (10.9b) is equivalent to (10.9a) and may essentially be viewed as also being a problem in the \(x\)-variable space. Moreover, under the condition of Theorem 9.3.1, if \( \mu \) is sufficiently large and if \( \bar{x} \) is an optimum for \( P \), then \( \bar{x} \) solves PP. Alternatively, as in Exercise 9.13, if \( \mu \) is sufficiently large and if \( \bar{x} \) satisfies the second-order sufficiency conditions for \( P \), then \( \bar{x} \) is a strict local minimum for \( PP \). In either case, \( \mu \) must be at least as large as the absolute value of any Lagrange multiplier associated with the constraints \( g(x) \leq 0 \) and \( h(x) = 0 \) in \( P \). Note that instead of using a single penalty parameter \( \mu \), we can employ a set of parameters \( \mu_1, \ldots, \mu_m, \mu_{m+\ell} \), one associated with each of the penalized constraints. Selecting some reasonably large values for these parameters (assuming a well-scaled problem), we can solve Problem PP; and if an infeasible solution results, then these parameters can be increased manually and the process repeated. We shall assume, however, that we have selected some suitably large, admissible value of a single penalty parameter \( \mu \). With this motivation, the algorithm PSLP seeks to solve Problem PP, using a box-step or hypercube first-order trust region approach, as introduced in Section 8.7.

Specifically, this approach proceeds as follows. Given a current iterate \( x_k \in X \) and a trust region or step-bound vector \( \Delta_k \in \mathbb{R}^n \), consider the following linearization of PP, given by (10.9a), where we have also imposed the given trust region step bound on the variation of \( x \) about \( x_k \).
Methods of Feasible Directions

\[ \text{LP}(x_k, \Delta_k): \text{Minimize } F_{EL_k}(x) = f(x_k) + \nabla f(x_k)^T(x-x_k) \]

\[ + \mu \left[ \sum_{i=1}^{m} \max\{0, g_i(x_k) + \nabla g_i(x_k)^T(x-x_k)\} \right] \]

\[ + \sum_{i=1}^{\ell} \left| h_i(x_k) + \nabla h_i(x_k)^T(x-x_k) \right| \]

subject to \( x \in X = \{x: Ax \leq b\} \)

\[ -\Delta_k \leq x-x_k \leq \Delta_k. \]

Similar to (10.9) and (10.10), this can be equivalently restated as the following linear programming problem, where we have also used the substitution \( x = x_k + d \) and have dropped the constant \( f(x_k) \) from the objective function:

\[ \text{LP}(x_k, \Delta_k): \text{Minimize } \nabla f(x_k)^T d + \mu \left[ \sum_{i=1}^{m} y_i + \sum_{i=1}^{\ell} (z_i^+ + z_i^-) \right] \]

subject to \( y_i \geq g_i(x_k) + \nabla g_i(x_k)^T d, \quad i = 1, \ldots, m \)

\( (z_i^+ - z_i^-) = h_i(x_k) + \nabla h_i(x_k)^T d \quad i = 1, \ldots, \ell \) (10.11b)

\( A(x_k + d) \leq b \)

\( -\Delta_k \leq d_i \leq \Delta_k \quad i = 1, \ldots, n \)

\( y \geq 0, z^+ \geq 0, z^- \geq 0. \)

The linear program \( \text{LP}(x_k, \Delta_k) \) given by (10.11b) is the direction-finding subproblem that yields an optimal solution \( d_k \), say, along with the accompanying values of \( y, z^+, \) and \( z^- \), which are given as follows, similar to (10.10):

\[ y_i = \max\{0, g_i(x_k) + \nabla g_i(x_k)^T d_k\}, \quad i = 1, \ldots, m \]

\[ z_i^+ = \max\{0, h_i(x_k) + \nabla h_i(x_k)^T d_k\}, \quad \text{(10.12)} \]

\[ z_i^- = \max\{0, -[h_i(x_k) + \nabla h_i(x_k)^T d_k]\}, \]

so that \( (z_i^+ + z_i^-) = \left| h_i(x_k) + \nabla h_i(x_k)^T d_k \right|, \quad i = 1, \ldots, \ell. \)

As with trust region methods described in Section 8.7, the decision whether to accept or to reject the new iterate \( x_k + d_k \) and the adjustment of the step bounds \( \Delta_k \) is made based on the ratio \( R_k \) of the actual decrease \( \Delta F_{EL_k} \) in the \( \ell_1 \) penalty function \( F_E \), and the decrease \( \Delta F_{EL_k} \) as predicted by its linearized version.
provided that the latter is nonzero. These quantities are given as follows from (10.8) and (10.11a):

\[ \Delta F_{E_k} = F_E(x_k) - F_E(x_k + d_k) \quad \Delta F_{EL_k} = F_{EL_k}(x_k) - F_{EL_k}(x_k + d_k). \]  

(10.13)

The principal concepts tying in the development presented thus far are encapsulated by the following result.

### 10.3.1 Theorem

Consider Problem P and the absolute value (\( \ell_1 \)) penalty function (10.8), where \( \mu \) is assumed to be large enough, as in Theorem 9.3.1.

a. If the conditions of Theorem 9.3.1 hold true and if \( \bar{x} \) solves Problem P, then \( \bar{x} \) also solves PP of Equation (10.9a). Alternatively, if \( \bar{x} \) is a regular point that satisfies the second-order sufficiency conditions for P, then \( \bar{x} \) is a strict local minimum for PP.

b. Consider Problem PP given by (10.9b), where \((y, z^+, z^-)\) are given by (10.10) for any \( x \in X \). If \( \bar{x} \) is a KKT solution for Problem P, then for \( \mu \) large enough, as in Theorem 9.3.1, \( \bar{x} \) is a KKT solution for Problem PP. Conversely, if \( \bar{x} \) is a KKT solution for PP and if \( \bar{x} \) is feasible to P, then \( \bar{x} \) is a KKT solution for P.

c. The solution \( d_k = 0 \) is optimal for \( LP(x_k, \Delta_k) \) defined by (10.11b) and (10.12) if and only if \( x_k \) is a KKT solution for PP.

d. The predicted decrease \( \Delta F_{EL_k} \) in the linearized penalty function, as given by (10.13), is nonnegative, and is zero if and only if \( d_k = 0 \) solves Problem \( LP(x_k, \Delta_k) \).

**Proof**

The proof for Part a is similar to that of Theorem 9.3.1 and of Exercise 9.13, and is left to the reader in Exercise 10.17. Next, consider Part b. The KKT conditions for P require a primal feasible solution \( \bar{x} \), along with Lagrange multipliers \( \bar{u}, \bar{v}, \) and \( \bar{w} \) satisfying

\[
\sum_{i=1}^{m} \bar{u}_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} \bar{v}_i \nabla h_i(\bar{x}) + A^T \bar{w} = -\nabla f(\bar{x})
\]

\[
\bar{u} \geq 0, \quad \bar{v} \text{ unrestricted,} \quad \bar{w} \geq 0
\]

(10.14)

\[
\bar{u}^T g(\bar{x}) = 0, \quad \bar{w}^T (A\bar{x} - b) = 0.
\]
Furthermore, $\bar{x}$ is a KKT point for PP, with $\bar{x} \in X$ and with $(y, z^+, z^-)$ given accordingly by (10.10), provided that there exist Lagrange multipliers $\bar{u}$, $\bar{v}$, and $\bar{w}$ satisfying

$$\sum_{i=1}^{m} \bar{u}_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} \bar{v}_i \nabla h_i(\bar{x}) + A' \bar{w} = -\nabla f(\bar{x})$$

(10.15a)

$$0 \leq \bar{u}_i \leq \mu, \ (\bar{u}_i - \mu)y_i = 0, \ \bar{u}_i[y_i - g_i(\bar{x})] = 0, \ i = 1, \ldots, m$$

(10.15b)

$$|\bar{v}_i| \leq \mu, \ z_i^+(\mu - \bar{v}_i) = 0, \ z_i^-(\mu + \bar{v}_i) = 0, \ i = 1, \ldots, \ell$$

(10.15c)

$$\bar{w}'(A\bar{x} - b) = 0, \ \bar{w} \geq 0.$$  

(10.15d)

Now let $\bar{x}$ be a KKT solution for Problem P, with Lagrange multipliers $\bar{u}$, $\bar{v}$, and $\bar{w}$ satisfying (10.14). Defining $(y, z^+, z^-)$ according to (10.10), we get $y = 0$, $z^+ = z^- = 0$; so for $\mu$ large enough, as in Theorem 9.3.1, $\bar{x}$ is a KKT solution for PP by (10.15). Conversely, let $\bar{x}$ be a KKT solution for PP and suppose that $x$ is feasible to Problem P. Then we again have $y = 0$, $z^+ = z^- = 0$ by (10.10); so, by (10.15) and (10.14), $\bar{x}$ is KKT solution for Problem P. This proves Part b.

Part c follows from Theorem 4.2.15, noting that LP$(x_k, A_k)$ represents a first-order linearization of PP at the point $x_k$ and that the step bounds $-\Delta_k \leq d \leq \Delta_k$ are nonbinding at $d_k = 0$, that is, at $x = x_k$.

Finally, consider Part d: Since $d_k$ minimizes LP$(x_k, A_k)$ in (10.11b), $x = x_k + d_k$ minimizes (10.11a); so since $x_k$ is feasible to (10.11a), we have that $F_{EL_k}(x_k) \geq F_{EL_k}(x_k + d_k)$, or that $\Delta F_{EL_k} \geq 0$. By the same token, this difference is zero if and only if $d_k = 0$ is optimal for LP$(x_k, A_k)$, and this completes the proof.

**Summary of the Penalty Successive Linear Programming (PPLP) Algorithm**

**Initialization** Put the iteration counter $k = 1$, and select a starting solution $x_k \in X$ feasible to the linear constraints, along with a step bound or trust region vector $\Delta_k > 0$ in $R^n$. Let $\Delta_{LB} > 0$ be some small lower bound tolerance on $\Delta_k$. (Sometimes, $\Delta_{LB} = 0$ is also used.) Additionally, select a suitable value of the penalty parameter $\mu$ (or values for penalty parameters $\mu_1, \ldots, \mu_m + \ell$, as discussed above). Choose values for the scalars $0 < \rho_0 < \rho_1 < \rho_2 < 1$ to be used in the trust region ratio test, and for the step bound adjustment multiplier $\beta \in (0, 1)$. (Typically, $\rho_0 = 10^{-6}$, $\rho_1 = 0.25$, $\rho_2 = 0.75$, and $\beta = 0.5$.)
**Step 1: Linear Programming Subproblem** Solve the linear program \( \text{LP}(x_k, \Delta_k) \) to obtain an optimum \( d_k \). Compute the actual and predicted decreases \( \Delta F_{E_k} \) and \( \Delta F_{Elk} \), respectively, in the penalty function as given by (10.13). If \( \Delta F_{Elk} = 0 \) (equivalently, by Theorem 10.3.1d, if \( d_k = 0 \)), then stop. Otherwise, compute the ratio \( R_k = \frac{\Delta F_{E_k}}{\Delta F_{Elk}} \). If \( R_k < \rho_0 \), then since \( \Delta F_{Elk} > 0 \) by Theorem 10.3.1d, the penalty function has either worsened or its improvement is insufficient. Hence, reject the current solution, shrink \( \Delta_k \) to \( \beta \Delta_k \), and repeat this step. (Zhang et al. [1985] show that within a finite number of such reductions, we will have \( R_k \geq \rho_0 \). Note that while \( R_k \) remains less than \( \rho_0 \), some components of \( \Delta_k \) may shrink below those of \( \Delta_{LB} \).) On the other hand, if \( R_k \geq \rho_0 \), proceed to Step 2.

**Step 2: New Iterate and Adjustment of Step Bounds** Let \( x_{k+1} = x_k + d_k \). If \( \rho_1 \leq R_k < \rho_2 \), then shrink \( \Delta_k \) to \( \Delta_{k+1} = \beta \Delta_k \), since the penalty function has not improved sufficiently. If \( \rho_1 \leq R_k \leq \rho_2 \), then retain \( \Delta_{k+1} = \Delta_k \). On the other hand, if \( R_k > \rho_2 \), amplify the trust region by letting \( \Delta_{k+1} = \Delta_k / \beta \). In all cases, replace \( \Delta_{k+1} \) by \( \max \{ \Delta_{k+1}, \Delta_{LB} \} \), where \( \max \{ \} \) is taken componentwise. Increment \( k \) by 1 and go to Step 1.

A few comments are in order at this point. First, note that the linear program (10.11b) is feasible and bounded (\( d = 0 \) is a feasible solution) and that it preserves any sparsity structure of the original problem. Second, if there are any variables that appear linearly in the objective function as well as in the constraints of \( P \), the corresponding step bounds for such variables can be taken as some arbitrarily large value \( M \) and can be retained at that value throughout the procedure. Third, when termination occurs at Step 1, then by Theorem 10.3.1, \( x_k \) is a KKT solution for \( PP \); and if \( x_k \) is feasible to \( P \), then it is also a KKT solution for \( P \). (Otherwise, the penalty parameters may need to be increased as discussed earlier.) Fourth, it can be shown that either the algorithm terminates finitely or else an infinite sequence \( \{ x_k \} \) is generated such that if the level set \( \{ x \in X : F_E(x) \leq F_E(x_1) \} \) is bounded, then \( \{ x_k \} \) has an accumulation point, and every such accumulation point is a KKT solution for Problem \( PP \). Finally, the stopping criterion of Step 1 is usually replaced by several practical termination criteria. For example, if the fractional change in the \( \ell_1 \) penalty function is less than a tolerance \( \varepsilon = 10^{-4} \) for some \( c = 3 \) consecutive iterations, or if the iterate is \( \varepsilon \)-feasible and either the KKT conditions are satisfied within an \( \varepsilon \)-tolerance, or if the fractional change in the objective function value for Problem \( P \) is less than \( \varepsilon \) for \( c \) consecutive iterations, the procedure can be terminated. Also, the amplification or reduction in the step bounds are often modified in implementations so as to treat deviations of \( R_k \).
from unity symmetrically. For example, at Step 2, if \( |1 - R_k| < 0.25 \), then all step bounds are amplified by dividing by \( \beta = 0.5 \); and if \( |1 - R_k| > 0.75 \), then all step bounds are reduced by multiplying by \( \beta \). In addition, if any variable that appears in a nonlinear term within the original problem remains at the same step bound for \( c (= 3) \) consecutive iterations, then its step bound is amplified by dividing by \( \beta \).

### 10.3.2 Example

Consider the problem:

\[
\begin{align*}
\text{Minimize } & \quad f(x) = 2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2 \\
\text{subject to } & \quad g_1(x) = 2x_1^2 - x_2 \leq 0 \\
& \quad x \in X = \{x = (x_1, x_2) : x_1 + 5x_2 \leq 5, x \geq 0\}.
\end{align*}
\]

Figure 10.13a provides a sketch for the graphical solution of this problem. Note that this problem has a "vertex" solution, and thus we might expect a rapid convergence behavior. Let us begin with the solution \( x_1 = (0, 1)^T \in X \) and use \( \mu = 10 \) (which can be verified to be sufficiently large—see Exercise 10.20). Let us also select \( \Delta_1 = (1, 1)^T, \Delta_{LB} = (10^{-6}, 10^{-6})^T, \rho_0 = 10^{-6}, \rho_1 = 0.25, \rho_2 = 0.75, \rho_3 = 0.5 \).

The linear program \( LP(x_1, \Delta_1) \) given by (10.11b) now needs to be solved, as, for example, by the simplex method. To illustrate the process graphically, consider the equivalent problem (10.11a). Noting that \( x_1 = (0, 1)^T, \mu = 10, f(x_1) = -4, \nabla f(x_1) = (-6, -2)^T, g_1(x_1) = -1, \) and \( \nabla g_1(x_1) = (0, -1)^T, \) we have

\[
F_{E14}(x) = -2 - 6x_1 - 2x_2 + 10 \max\{0, -x_2\}. \tag{10.16}
\]

![Figure 10.13 Solution to Example 10.3.2.](image-url)
The solution of \( LP(x_1, A_1) \) via (10.11a) is depicted in Figure 10.13b. The optimum solution is \( x = (1, 4/5)^T \), so that \( d_1 = (1, 4/5)^T - (0, 1)^T = (1, -1/5)^T \) solves (10.11b). From (10.13), using (10.8) and (10.16) along with \( x_1 = (0, 1)^T \) and \( x_1 + d_1 = (1, 4/5)^T \), we get \( \Delta F_{E_k} = -4.88 \) and \( \Delta F_{EL_k} = 28/5 \). Hence, the penalty function has worsened, so we reduce the step bounds at Step 1 itself and repeat this step with the same point \( x_1 \) and with the revised \( A_1 = (0.5, 0.5)^T \).

The revised step bound box is shown dashed in Figure 10.13b. The corresponding optimal solution is \( x = (0.5, 0.9)^T \), which corresponds to the optimum \( d_1 = (0.5, 0.9)^T - (0, 1)^T = (0.5, -0.1)^T \) for the problem (10.11b). From (10.13), using (10.8) and (10.16) along with \( x_1 = (0, 1)^T \) and \( x_1 + d_1 = (0.5, 0.9)^T \), we get \( \Delta F_{E_k} = 2.18 \) and \( \Delta F_{EL_k} = 2.8 \), which gives \( R_k = 2.18/2.8 = 0.7786 \). We therefore accept this solution as the new iterate \( x_2 = (0.5, 0.9)^T \), and since \( R_k > \rho_2 = 0.75 \), we amplify the trust region by letting \( \Delta_2 = \Delta_1/\beta = (1, 1)^T \).

We now ask the reader (see Exercise 10.20) to continue this process until a suitable termination criterion is satisfied, as discussed above.

### 10.4 Successive Quadratic Programming or Projected Lagrangian Approach

We have seen that Zoutendijk’s algorithm as well as Topkis and Veinott’s modification of this procedure are prone to zigzagging and slow convergence behavior because of the first-order approximations employed. The SLP approach enjoys a quadratic rate of convergence if the optimum occurs at a vertex of the feasible region, because then the method begins to imitate Newton’s method applied to the active constraints. However, for nonvertex solutions, this method again, being essentially a first-order approximation procedure, can succumb to a slow convergence process. To alleviate this behavior, we can employ second-order approximations and derive a successive quadratic programming approach (SQP).

SQP methods, also known as sequential, or recursive, quadratic programming approaches, employ Newton’s method (or quasi-Newton methods) to directly solve the KKT conditions for the original problem. As a result, the accompanying subproblem turns out to be the minimization of a quadratic approximation to the Lagrangian function optimized over a linear approximation to the constraints. Hence, this type of process is also known as a projected Lagrangian, or Lagrange-Newton, approach. By its nature this method produces both primal and dual (Lagrange multiplier) solutions.

To present the concept of this method, consider the equality-constrained nonlinear problem, where \( x \in R^n \), and all functions are assumed to be continuously twice differentiable.
Methods of Feasible Directions

P: Minimize $f(x)$
subject to $h_i(x) = 0, \quad i = 1, \ldots, \ell$. \hspace{2cm} (10.17)

The extension for including inequality constraints is motivated by the following analysis for the equality-constrained case and is considered subsequently.

The KKT optimality conditions for Problem P require a primal solution $x \in \mathbb{R}^n$ and a Lagrange multiplier vector $v \in \mathbb{R}^\ell$ such that

$$
\nabla f(x) + \sum_{i=1}^\ell v_i \nabla h_i(x) = 0
$$

$$
h_i(x) = 0, \quad i = 1, \ldots, \ell. \hspace{2cm} (10.18)
$$

Let us write this system of equations more compactly as $W(x, v) = 0$. We now use the Newton-Raphson method to solve (10.18) or, equivalently, use Newton’s method to minimize a function for which (10.18) represents the first-order condition that equates the gradient to zero. Hence, given an iterate $(x_k, v_k)$, we solve the first-order approximation

$$
W(x_k, v_k) + \nabla W(x_k, v_k) \begin{bmatrix} x-x_k \\ v-v_k \end{bmatrix} = 0 \hspace{2cm} (10.19)
$$

to the given system to determine the next iterate $(x, v) = (x_{k+1}, v_{k+1})$, where $\nabla W$ denotes the Jacobian of $W$. Defining $\nabla^2 L(x_k) = \nabla^2 f(x_k) + \sum_{i=1}^\ell v_{ki}$, $\nabla^2 h_i(x_k)$ to be the usual Hessian of the Lagrangian at $x_k$ with the Lagrange multiplier vector $v_k$, and letting $\nabla h$ denote the Jacobian of $h$ comprised of rows $\nabla h_i(x)^T$ for $i = 1, ..., \ell$, we have

$$
\nabla W(x_k, v_k) = \begin{bmatrix} \nabla^2 L(x_k) & \nabla h(x_k)^T \\ \nabla h(x_k) & 0 \end{bmatrix}. \hspace{2cm} (10.20)
$$

Using (10.18) and (10.20), we can rewrite (10.19) as

$$
\nabla^2 L(x_k)(x-x_k) + \nabla h(x_k)^T(v-v_k) = -\nabla f(x_k) - \nabla h(x_k)^T v_k
$$

$$
\nabla h(x_k)(x-x_k) = -h(x_k). \hspace{2cm} (10.21)
$$

Substituting $d = x - x_k$, this in turn can be rewritten as

$$
\nabla^2 L(x_k)d + \nabla h(x_k)^Tv = -\nabla f(x_k)
$$

$$
\nabla h(x_k)d = -h(x_k). \hspace{2cm} (10.21)
$$
We can now solve for \((d, v) = (d_k, v_{k+1})\), say, using this system, if a solution exists. (See the convergence analysis below and Exercise 10.22.) Setting \(x_{k+1} = x_k + d_k\), we then increment \(k\) by 1 and repeat this process until \(d = 0\) happens to solve \((10.21)\). When this occurs, if at all, noting \((10.18)\), we shall have found a KKT solution for Problem \(P\).

Now, instead of adopting the foregoing process to find any KKT solution for \(P\), we can, instead, employ a quadratic minimization subproblem whose optimality conditions duplicate \((10.21)\) but which might tend to drive the process toward beneficial KKT solutions. Such a quadratic program is stated below, where the constant term \(f(x_k)\) has been inserted into the objective function for insight and convenience.

\[
\text{QP}(x_k, v_k): \quad \text{Minimize} \quad f(x_k) + \nabla f(x_k)'d + \frac{1}{2}d'\nabla^2 L(x_k)d
\]

subject to \(h_i(x_k) + \nabla h_i(x_k)'d = 0, \quad i = 1, \ldots, \ell.\) (10.22)

Several comments regarding the linearly constrained quadratic subproblem \(\text{QP}(x_k, v_k)\), abbreviated \(\text{QP}\) whenever unambiguous, are in order at this point. First, note that an optimum to \(\text{QP}\), if it exists, is a KKT point for \(\text{QP}\) and satisfies \((10.21)\), where \(v\) is the set of Lagrange multipliers associated with the constraints of \(\text{QP}\). However, the minimization process of \(\text{QP}\) drives the solution toward a desirable KKT point satisfying \((10.21)\) whenever alternatives exist. Second, observe that by the foregoing derivation, the objective function of \(\text{QP}\) represents not just a quadratic approximation for \(f(x)\) but also incorporates an additional term \((1/2)\sum_{i=1}^\ell v_{ki}d'\nabla^2 h_i(x_k)d\) to represent the curvature of the constraints. In fact, defining the Lagrangian function \(L(x) = f(x) + \sum_{i=1}^\ell v_{ki}h_i(x)\), the objective function of \(\text{QP}(x_k, v_k)\) can be written alternatively as follows, noting the constraints:

\[
\text{Minimize} \quad L(x_k) + \nabla_x L(x_k)'d + \frac{1}{2}d'\nabla^2 L(x_k)d. \quad (10.23)
\]

Observe that \((10.23)\) represents a second-order Taylor series approximation for the Lagrangian function \(L\). In particular, this supports the quadratic convergence rate behavior in the presence of nonlinear constraints (see also Exercise 10.24). Third, note that the constraints of \(\text{QP}\) represent a first-order linearization at the current point \(x_k\). Fourth, observe that \(\text{QP}\) might be unbounded or infeasible, whereas \(P\) is not. Although the first of these unfavorable events can be managed by bounding the variation in \(d\), for instance, the second is more disconcerting. For example, if we have a constraint \(x_1^2 + x_2^2 = 1\) and we linearize this at the origin, we obtain an inconsistent restriction requiring that \(-1 = 0\). We later present a variant of the scheme above that overcomes this difficulty (see also...
Exercise 10.26). Notwithstanding this problem, and assuming a well-behaved QP subproblem, we are now ready to state a rudimentary SQP algorithm.

**Rudimentary SQP Algorithm (RSQP)**

*Initialization* Put the iteration counter $k = 1$ and select a (suitable) starting primal-dual solution $(x_k, v_k)$.

*Main Step* Solve the quadratic subproblem $QP(x_k, v_k)$ to obtain a solution $d_k$ along with a vector of Lagrange multipliers $v_{k+1}$. If $d_k = 0$, then from (10.21), $(x_k, v_{k+1})$ satisfies the KKT conditions (10.18) for Problem P; stop. Otherwise, put $x_{k+1} = x_k + d_k$, increment $k$ by 1, and repeat the Main Step.

**Convergence Rate Analysis**

Under appropriate conditions, we can argue a quadratic convergence behavior for the foregoing algorithm. Specifically, suppose that $\bar{x}$ is a regular KKT solution for Problem P which together with a set of Lagrange multipliers $\bar{v}$, satisfies the second-order sufficiency conditions of Theorem 4.4.2. Then $\nabla W(\bar{x}, \bar{v}) = \bar{V}W$, defined by (10.20), say, is nonsingular. To see this, let us show that the system

\[
\nabla W(\bar{x}, \bar{v}) \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = 0
\]

has the unique solution given by $(d'_1, d'_2) = 0$. Consider any solution $(d'_1, d'_2)$. Since $\bar{x}$ is a regular solution, $\nabla h(\bar{x})'$ has full column rank; so if $d_1 = 0$, then $d_2 = 0$ as well. If $d_1 \neq 0$, since $\nabla h(\bar{x})d_1 = 0$, we have by the second-order sufficiency conditions that $d_1'\nabla^2 L(\bar{x})d_1 > 0$. However, since $\nabla^2 L(\bar{x})d_1 + \nabla h(\bar{x})d_2 = 0$, we have that $d_1'\nabla^2 L(\bar{x})d_1 = -d_2'\nabla h(\bar{x})d_1 = 0$, a contradiction. Hence, $\nabla W$ is nonsingular; and thus for $(x_k, v_k)$ sufficiently close to $(\bar{x}, \bar{v})$, $\nabla W(x_k, v_k)$ is nonsingular. Therefore, the system (10.21), and thus Problem $QP(x_k, v_k)$, has a well-defined (unique) solution. Consequently, in the spirit of Theorem 8.6.5, when $(x_k, v_k)$ is sufficiently close to $(\bar{x}, \bar{v})$, a quadratic rate of convergence to $(\bar{x}, \bar{v})$ is obtained.

Actually, the closeness of $x_k$ alone to $\bar{x}$ is sufficient to establish convergence. It can be shown (see the Notes and References section) that if $x_1$ is sufficiently close to $\bar{x}$ and if $\nabla W(x_1, v_1)$ is nonsingular, the algorithm RSQP converges quadratically to $(\bar{x}, \bar{v})$. In this respect, the Lagrange multipliers $v$, appearing only in the second-order term in QP, do not play as important a role as
they do in augmented Lagrangian (ALAG) penalty methods, for example, and inaccuracies in their estimation can be tolerated more flexibly.

**Extension to Include Inequality Constraints**

We now consider the inclusion of inequality constraints \( g_i(x) \leq 0, \ i = 1, \ldots, m \), in Problem P, where \( g_i \) are continuously twice differentiable for \( i = 1, \ldots, m \). This revised problem is restated below.

\[
P: \text{Minimize } f(x) \quad \text{subject to } \quad g_i(x) \leq 0, \ i = 1, \ldots, m \tag{10.24}
\]

\[
h_i(x) = 0, \ i = 1, \ldots, \ell.
\]

For this instance, given an iterate \((x_k, u_k, v_k)\), where \( u_k \geq 0 \) and \( v_k \) are, respectively, the Lagrange multiplier estimates for the inequality and the equality constraints, we consider the following quadratic programming subproblem as a direct extension of (10.22):

\[
\text{QP}(x_k, u_k, v_k): \text{Minimize } f(x_k) + \nabla f(x_k)'d + \frac{1}{2} d'\nabla^2 L(x_k)d
\]

subject to \( g_i(x_k) + \nabla g_i(x_k)'d \leq 0, \ i = 1, \ldots, m \tag{10.25} \)

\[
h_i(x_k) + \nabla h_i(x_k)'d = 0, \ i = 1, \ldots, \ell
\]

where \( \nabla^2 L(x_k) = \nabla^2 f(x_k) + \sum_{i=1}^m u_k g_i(x_k) + \sum_{i=1}^\ell v_k h_i(x_k) \). Note that the KKT conditions for this problem require that in addition to primal feasibility, we find Lagrange multipliers \( u \) and \( v \) such that

\[
\nabla f(x_k) + \nabla^2 L(x_k)d + \sum_{i=1}^m u_i \nabla g_i(x_k) + \sum_{i=1}^\ell v_i \nabla h_i(x_k) = 0 \tag{10.26a}
\]

\[
u_i[g_i(x_k) + \nabla g_i(x_k)'d] = 0, \ i = 1, \ldots, m \tag{10.26b}
\]

\[
u \geq 0, \ v \text{ unrestricted.} \tag{10.26c}
\]

Hence, if \( d_k \) solves QP\((x_k, u_k, v_k)\) with Lagrange multipliers \( u_{k+1} \) and \( v_{k+1} \), and if \( d_k = 0 \), then \( x_k \) along with \((u_{k+1}, v_{k+1})\) yields a KKT solution for the original Problem P. Otherwise, we set \( x_{k+1} = x_k + d_k \) as before, increment \( k \) by 1, and repeat the process. In a similar manner, it can be shown that if \( \bar{x} \) is a regular KKT solution which, together with \((\bar{u}, \bar{v})\), satisfies the second-order sufficiency conditions, and if \((x_k, u_k, v_k)\) is initialized sufficiently close to \((\bar{x}, \bar{u}, \bar{v})\), the foregoing iterative process will converge quadratically to \((\bar{x}, \bar{u}, \bar{v})\).
Quasi-Newton Approximations

One disadvantage of the SQP method discussed thus far is that we require second-order derivatives to be calculated and, besides, that $\nabla^2 L(x_k)$ might not be positive definite. This can be overcome by employing quasi-Newton positive definite approximations for $\nabla^2 L$. For example, given a positive definite approximation $B_k$ for $\nabla^2 L(x_k)$ in the algorithm RSQP described above, we can solve the system (10.21) with $\nabla^2 L(x_k)$ replaced by $B_k$, to obtain the unique solution $d_k$ and $v_{k+1}$ and then set $x_{k+1} = x_k + d_k$. This is equivalent to the iterative step given by

$$
\begin{bmatrix}
  x_{k+1} \\
  v_{k+1}
\end{bmatrix}
= \begin{bmatrix}
  x_k \\
  v_k
\end{bmatrix} - \begin{bmatrix}
  B_k & \nabla h(x_k)' \\
  \nabla h(x_k) & 0
\end{bmatrix}^{-1} \begin{bmatrix}
  \nabla L(x_k) \\
  h(x_k)
\end{bmatrix},
$$

where $\nabla L(x_k) = \nabla f(x_k) + \nabla h(x_k)'v_k$. Then, adopting the popular BFGS update for the Hessian as defined by (10.88.63), we can compute

$$
B_{k+1} = B_k + \frac{q_k q_k'}{p_k p_k} - \frac{B_k p_k p_k' B_k}{p_k' B_k p_k},
$$

where

$$
p_k = x_{k+1} - x_k, \quad q_k = \nabla L'(x_{k+1}) - \nabla L'(x_k),
$$

and where

$$
\nabla L'(x) = \nabla f(x) + \sum_{i=1}^{\ell} \nu_{(k+1)i} \nabla h_i(x).
$$

It can be shown that this modification of the rudimentary process, similar to the quasi-Newton modification of Newton's algorithm, converges superlinearly when initialized sufficiently close to a solution $(\bar{x}, \bar{v})$ that satisfies the foregoing regularity and second-order sufficiency conditions. However, this superlinear convergence rate is based strongly on the use of unit step sizes.

Globally Convergent Variant Using the $l_1$ Penalty as a Merit Function

A principal disadvantage of the SQP method described thus far is that convergence is guaranteed only when the algorithm is initialized sufficiently close to a desirable solution, whereas, in practice, this condition is usually difficult to realize. To remedy this situation and to ensure global convergence, we introduce the idea of a merit function. This is a function that along with the objective function is minimized simultaneously at the solution of the problem, but one that also serves as a descent function, guiding the iterates and providing a measure of progress. Preferably, it should be easy to evaluate this function,
and it should not impair the convergence rate of the algorithm. We describe the use of the popular $\ell_1$, or absolute value, penalty function (9.8), restated below, as a merit function for Problem P given in (10.24):

$$F_E(x) = f(x) + \mu \left[ \sum_{i=1}^{m} \max\{0, g_i(x)\} + \sum_{i=1}^{\ell} |h_i(x)| \right]. \tag{10.28}$$

The following lemma establishes the role of $F_E$ as a merit function. The Notes and References section points out other quadratic and ALAG penalty functions that can be used as merit functions in a similar context.

### 10.4.1 Lemma

Given an iterate $x_k$, consider the quadratic subproblem QP given by (10.25), where $V^2 L(x_k)$ is replaced by any positive definite approximation $B_k$. Let $d$ solve this problem with Lagrange multipliers $u$ and $v$ associated with the inequality and the equality constraints, respectively. If $d \neq 0$, and if $\mu \geq \max\{u_1, u_m, |v_1|, \ldots, |v_\ell|\}$, then $d$ is a descent direction at $x = x_k$ for the $\ell_1$ penalty function $F_E$ given by (10.28).

**Proof**

Using the primal feasibility, the dual feasibility, and the complementary slackness conditions (10.25), (10.26a), and (10.26b) for QP, we have

$$\nabla f(x_k)'d = -d' B_k d - \sum_{i=1}^{m} u_i \nabla g_i(x_k)'d - \sum_{i=1}^{\ell} v_i \nabla h_i(x_k)'d$$

$$= -d' B_k d + \sum_{i=1}^{m} u_i g_i(x_k) + \sum_{i=1}^{\ell} v_i h_i(x_k)$$

$$\leq -d' B_k d + \sum_{i=1}^{m} u_i \max\{0, g_i(x_k)\} + \sum_{i=1}^{\ell} |v_i||h_i(x_k)|$$

$$\leq -d' B_k d + \mu \left[ \sum_{i=1}^{m} \max\{0, g_i(x_k)\} + \sum_{i=1}^{\ell} |h_i(x_k)| \right]. \tag{10.29}$$

Now, from (10.28), we have that for a step length $\lambda \geq 0$,

$$F_E(x_k) - F_E(x_k + \lambda d) = [f(x_k) - f(x_k + \lambda d)]$$

$$+ \mu \left[ \sum_{i=1}^{m} \max\{0, g_i(x_k)\} - \max\{0, g_i(x_k + \lambda d)\} \right]$$

$$+ \sum_{i=1}^{\ell} [h_i(x_k) - h_i(x_k + \lambda d)]. \tag{10.30}$$
Letting \( O_i(\lambda) \) denote an appropriate function that approaches zero as \( \lambda \to 0 \), for \( i = 1, \ldots, m + \ell \), we have, for \( \lambda > 0 \) and sufficiently small,

\[
f(x_k + \lambda d) = f(x_k) + \lambda \nabla f(x_k)'d + \lambda O_0(\lambda). \tag{10.31a}
\]

Also, \( g_i(x_k + \lambda d) = g_i(x_k) + \lambda \nabla g_i(x_k)'d + \lambda O_i(\lambda) \leq g_i(x_k) - \lambda g_i(x_k) + \lambda O_i(\lambda) \) from (10.25). Hence,

\[
\max\{0, g_i(x_k + \lambda d)\} \leq (1 - \lambda) \max\{0, g_i(x_k)\} + \lambda|O_i(\lambda)|. \tag{10.31b}
\]

Similarly, from (10.25),

\[
h_i(x_k + \lambda d) = h_i(x_k) + \lambda \nabla h_i(x_k) + \lambda O_{m+i}(\lambda) = (1 - \lambda)h_i(x_k) + \lambda O_{m+i}(\lambda),
\]

and hence

\[
|h_i(x_k + \lambda d)| \leq (1 - \lambda)|h_i(x_k)| + \lambda|O_{m+i}(\lambda)|. \tag{10.31c}
\]

Using (10.31) in (10.30), we obtain for \( \lambda \geq 0 \) and sufficiently small, that \( F_E(x_k) - F_E(x_k + \lambda d) \geq \lambda \{ -\nabla f(x_k)'d + \mu \{ \sum_{i=1}^{m} \max\{0, g_i(x_k)\} + \sum_{i=1}^{\ell} |h_i(x_k)| \} + O(\lambda) \} \}, \)

where \( O(\lambda) \to 0 \) as \( \lambda \to 0 \). Hence, by (10.29), this gives \( F_E(x_k) - F_E(x_k + \lambda d) \geq \lambda d'B_kd + O(\lambda) > 0 \) for all \( \lambda \in (0, \delta) \) for some \( \delta > 0 \) by the positive definiteness of \( B_k \), and this completes the proof.

Lemma 10.4.1 exhibits flexibility in the choice of \( B_k \) for the resulting direction to be a descent direction for the exact penalty function. This matrix needs to be positive definite and may be updated by using any quasi-Newton strategy such as an extension of (10.27), or may even be held constant throughout the algorithm. This descent feature enables us to obtain a globally convergent algorithm under mild assumptions, as shown below.

**Summary of the Merit Function SQP Algorithm (MSQP)**

**Initialization** Put the iteration counter at \( k = 1 \) and select a (suitable) starting solution \( x_k \). Also, select a positive definite approximation \( B_k \) to the Hessian \( \nabla^2 L(x_k) \) defined with respect to some Lagrange multipliers \( u_k \geq 0 \) and \( v_k \) associated with the inequality and the equality constraints, respectively, of Problem (10.24). [Note that \( B_k \) might be arbitrary and need not necessarily bear any relationship to \( \nabla^2 L(x_k) \), although this is desirable.]

**Main Step** Solve the quadratic programming subproblem QP given by (10.25) with \( \nabla^2 L(x_k) \) replaced by \( B_k \) and obtain a solution \( d_k \) along with Lagrange multipliers \( (u_{k+1}, v_{k+1}) \). If \( d_k = 0 \), then stop with \( x_k \) as a KKT
solution for Problem P of (10.24), having Lagrange multipliers \((u_{k+1}, v_{k+1})\). Otherwise, find \(x_{k+1} = x_k + \lambda_k d_k\), where \(\lambda_k\) minimizes \(F_E(x_k + \lambda d_k)\) over \(\lambda \in \mathbb{R}, \lambda \geq 0\). Update \(B_k\) to a positive definite matrix \(B_{k+1}\) [which might be \(B_k\), itself, or \(\nabla^2 L(x_{k+1})\) defined with respect to \((u_{k+1}, v_{k+1})\), or some approximation thereof that is updated according to a quasi-Newton scheme]. Increment \(k\) by 1 and repeat the Main Step.

The reader may note that the line search above is to be performed with respect to a nondifferentiable function, which obviates the use of the techniques in Sections 8.2 and 8.3, including the popular curve-fitting approaches. Below we sketch the proof of convergence for algorithm MSQP. In Exercise 10.27, we ask the reader to provide a detailed argument.

10.4.2 Theorem

Algorithm MSQP either terminates finitely with a KKT solution to Problem P defined in (10.24), or else an infinite sequence of iterates \(\{x_k\}\) is generated. In the latter case, assume that \(\{x_k\} \subseteq X\), a compact subset of \(\mathbb{R}^n\), and that for any point \(x \in X\) and any positive definite matrix \(B\), the quadratic programming subproblem QP (with \(\nabla^2 L\) replaced by \(B\)) has a unique solution \(d\) (and so this problem is feasible), and has unique Lagrange multipliers \(u\) and \(v\) satisfying \(\mu \geq \max \{u_1, \ldots, u_m, |v_1|, \ldots, |v_r|\}\), where \(\mu\) is the penalty parameter for \(F_E\) defined in (10.28). Furthermore, assume that the accompanying sequence \(\{B_k\}\) of positive definite matrices generated lies in a compact subspace, with all accumulation points being positive definite (or with \(\{B_k^{-1}\}\) also being bounded). Then every accumulation point of \(\{x_k\}\) is a KKT solution for P.

Proof

Let the solution set \(\Omega\) be composed of all points \(x\) such that the corresponding subproblem QP produces \(d = 0\) at optimality. Note from (10.26) that given any positive definite matrix \(B\), \(x\) is a KKT solution for P if and only if \(d = 0\) is optimal for QP; that is, \(x \in \Omega\). Now the algorithm MSQP can be viewed as a map \(UMD\), where \(D\) is the direction-finding map that determines the direction \(d_k\) via the subproblem QP defined with respect to \(x_k\) and \(B_k\), \(M\) is the usual line search map, and \(U\) is a map that updates \(B_k\) to \(B_{k+1}\). Since the optimality conditions of QP are continuous in the data, the output of QP can readily be seen to be a continuous function of the input. By Theorem 8.4.1, the line search map \(M\) is also closed, since \(F_E\) is continuous. Since the conditions of Theorem 7.3.2 hold true, \(MD\) is therefore closed. Moreover, by Lemma 10.4.1, if \(x_k \notin \Omega\), then \(F_E(x_{k+1}) < F_E(x_k)\), thus providing a strict descent function. Since the map \(U\) does not disturb this descent feature; and since \(\{x_k\}\)
and \( \{ B_k \} \) are contained within compact sets, with any accumulation point of \( B_k \) being positive definite, the argument of Theorem 7.3.4 holds true. This completes the proof.

### 10.4.3 Example

To illustrate algorithms RSQP and MSQP, consider the following problem:

Minimize \[ 2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2 \]
subject to \[ g_1(x) = 2x_1^2 - x_2 \leq 0 \]
\[ g_2(x) = x_1 + 5x_2 - 5 \leq 0 \]
\[ g_3(x) = -x_1 \leq 0 \]
\[ g_4(x) = -x_2 \leq 0 \]

A graphical solution of this problem appears in Figure 10.13. Following Example 10.3.2, let us use \( \mu = 10 \) in the \( \ell _1 \) penalty merit function \( F_E \) defined by (10.28). Let us also use \( B_k = \nabla ^2 L (x_k) \) itself, and begin with \( x_1 = (0, 1)^t \) and with Lagrange multipliers \( u_1 = (0, 0, 0, 0)^t \). Hence, we have \( f(x_1) = -4 = F_E (x_1) \), since \( x_1 \) happens to be feasible. Also, \( g_1(x_1) = -1, g_2(x_1) = 0, g_3(x_1) = 0, \) and \( g_4(x_1) = -1 \). The function gradients are \( \nabla f(x_1) = (-6, -2)^t, \nabla g_1(x_1) = (0, 1, 5)^t, \nabla g_2(x_1) = (-1, 0)^t, \) and \( \nabla g_4(x_1) = (0, -1)^t \). The Hessian of the Lagrangian is

\[ \nabla ^2 L (x_1) = \nabla ^2 f(x_1) = \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix}. \]

Accordingly, the quadratic programming subproblem QP defined in (10.25) is as follows:

QP: Minimize \[ -6d_1 - 2d_2 + \frac{1}{2}[4d_1^2 + 4d_2^2 - 4d_1d_2] \]
subject to \[ -1 - d_2 \leq 0, \]
\[ d_1 + 5d_2 \leq 0 \]
\[ -d_1 \leq 0, \]
\[ -1 - d_2 \leq 0. \]

Figure 10.14 depicts the graphical solution of this problem. At optimality, only the second constraint of QP is binding. Hence, the KKT system gives

\[ 4d_1 - 2d_2 - 6 + u_2 = 0, \]
\[ 4d_2 - 2d_1 - 2 + 5u_2 = 0, \]
\[ d_1 + 5d_2 = 0. \]

Solving, we obtain \( d_1 = (35/31, -7/31)^t \) and \( u_2 = (0, 1.032258, 0, 0) \) as the primal and dual optimal solutions, respectively, to QP.
Now, for algorithm RSQP, we would take a unit step to obtain $x_2 = x_1 + d_1 = (1.1290322, 0.7741936)^T$. This completes one iteration. We ask the reader in Exercise 10.25 to continue this process and to examine its convergence behavior.

On the other hand, for algorithm MSQP, we need to perform a line search, minimizing $F_E$ from $x_1$ along the direction $d_1$. This line search problem is, from (10.32),

$$
\begin{align*}
\text{Minimize} \quad & F_E(x_1 + \lambda d_1) \\
= & [3.1612897 \lambda^2 - 6.3225804 \lambda - 4] \\
+ & 10[\max\{0, 2.5494274 \lambda^2 + 0.2258064 \lambda - 1\}] \\
+ & \max\{0, 0\} + \max\{0, -1.1290322\lambda\} \\
+ & \max\{0, -1+0.2258064\lambda\}.
\end{align*}
$$

Using the golden section method, for example, we find the step length $\lambda_1 = 0.5835726$. [Note that the unconstrained minimum of $f(x_1 + \lambda d_1)$ occurs at $\lambda = 1$; but beyond $\lambda = \lambda_1$, the first max{0, ·} term starts to become positive and increases the value of $F_E$, hence giving $\lambda_1$ as the desired step size.] This produces the new iterate $x_2 = x_1 + \lambda_1 d_1 = (0.6588722, 0.8682256)^T$. Observe that because the direction $d_1$ that was generated happened to be leading toward
the optimum for P, the minimization of the exact $\ell_1$ penalty function (with $\mu$ sufficiently large) produced this optimum. We ask the reader in Exercise 10.25 to verify the optimality of $x_2$ via the corresponding quadratic programming subproblem.

**Maratos Effect**

Consider the equality-constrained Problem P defined in (10.17). [A similar phenomenon holds true for Problem (10.24).] Note that the rudimentary SQP algorithm adopts a unit step size and converges quadratically when $(x_k, v_k)$ is initialized close to a regular solution $(\bar{x}, \bar{v})$ satisfying the second-order sufficiency conditions. The merit function-based algorithm, however, performs a line search at each iteration to minimize the exact penalty function $F_E$ of (10.28), given that the conditions of Lemma 10.4.1 hold true. Assuming all of the foregoing conditions, one might think that when $(x_k, v_k)$ is sufficiently close to $(\bar{x}, \bar{v})$, a unit step size would decrease the value of $F_E$. This statement is incorrect, and its violation is known as the Maratos effect, after N. Maratos, who discovered this in relation to Powell’s algorithm in 1978.

**10.4.4 Example (Maratos Effect)**

Consider the following example discussed in Powell (1986):

\[
\begin{align*}
\text{Minimize} & \quad f(x) = -x_1 + 2(x_1^2 + x_2^2 - 1) \\
\text{subject to} & \quad h(x) = x_1^2 + x_2^2 - 1 = 0.
\end{align*}
\]

Clearly, the optimum occurs at $\bar{x} = (1, 0)'$. The Lagrange multiplier at this solution is readily obtained from the KKT conditions to be $\bar{\nu} = -3/2$, so $\nabla^2 L(\bar{x}) = \nabla^2 f(\bar{x}) + \bar{\nu} \nabla^2 h(\bar{x}) = I$. Let us take the approximations $B_k$ to be equal to $I$ throughout the algorithm.

Now let us select $x_k$ to be sufficiently close to $\bar{x}$, but lying on the unit ball defining the constraint. Hence, we can let $x_k = (\cos \theta, \sin \theta)'$, where $|\theta|$ is small. The quadratic program (10.22) is given by:

\[
\begin{align*}
\text{Minimize} & \quad f(x_k) + (-1 + 4 \cos \theta)d_1 + (4 \sin \theta)d_2 + \frac{1}{2} (d_1^2 + d_2^2) \\
\text{subject to} & \quad 2 \cos \theta d_1 + 2 \sin \theta d_2 = 0
\end{align*}
\]

or, equivalently:

\[
\begin{align*}
\text{Minimize} & \quad \{f(x_k) - d_1 + \frac{1}{2} (d_1^2 + d_2^2) : \cos \theta d_1 + \sin \theta d_2 = 0\}.
\end{align*}
\]
Writing the KKT conditions for this problem and solving, we readily obtain the optimal solution \( \mathbf{d}_k = (\sin^2 \theta, -\sin \theta \cos \theta)' \). Hence, \( \mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k = (\cos \theta + \sin^2 \theta, \sin \theta - \sin \theta \cos \theta)' \). Note that \( \|\mathbf{x}_k - \mathbf{x}\|^2 = \sqrt{2(1 - \cos \theta)} = \theta \), adopting a second-order Taylor series approximation while, similarly, \( \|(\mathbf{x}_k + \mathbf{d}_k) - \mathbf{x}\| = \theta^2/2 \), thereby attesting to the rapid convergence behavior. However, it is readily verified that \( f(\mathbf{x}_k + \mathbf{d}_k) = -\cos \theta + \sin^2 \theta \) while \( f(\mathbf{x}_k) = -\cos \theta \) and also that \( h(\mathbf{x}_k + \mathbf{d}_k) = 2\sin \theta \) while \( h(\mathbf{x}_k) = 0 \). Hence, although a unit step makes \( \|x_k + d_k - x\| \) considerably smaller than \( \|x_k - x\| \), it results in an increase in both \( f \) and in the constraint violation, and therefore would increase the value of \( F_E \) for any \( \mu \geq 0 \), or for that matter, it would increase the value of any merit function.

Several suggestions have been proposed for overcoming the Maratos effect based on tolerating an increase in both \( f \) and the constraint violations, or recalculating the step length after correcting for second-order effects, or altering the search direction via modifications in second-order approximations to the objective and the constraint functions. We direct the reader to the Notes and References section for further reading on this subject.

Using the \( \ell_1 \) Penalty in the QP Subproblem: \( \ell_1 \) SQP Approach

In Section 10.3 we presented a superior penalty-based SLP algorithm that adopts trust region concepts and affords a robust and efficient scheme. A similar procedure has been proposed by Fletcher [1981] in the SQP framework, which exhibits a relatively superior computational behavior. Here, given an iterate \( \mathbf{x}_k \) and a positive definite approximation \( \mathbf{B}_k \) to the Hessian of the Lagrangian function, analogous to (10.11a), this procedure solves the following quadratic subproblem:

\[
\text{QP: Minimize } \left[ f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)' \mathbf{d} + \frac{1}{2} \mathbf{d}' \mathbf{B}_k \mathbf{d} \right]
\]

\[
+ \mu \left[ \sum_{i=1}^{m} \max \{0, g_i(\mathbf{x}_k) + \nabla g_i(\mathbf{x}_k)' \mathbf{d}\} \right]
\]

\[
+ \sum_{i=1}^{l} |h_i(\mathbf{x}_k) + \nabla h_i(\mathbf{x}_k)' \mathbf{d}| \right]
\]

subject to \( -\Delta_k \leq \mathbf{d} \leq \Delta_k \),

where \( \Delta_k \) is a trust region step bound, and as before, \( \mu \) is a suitably large penalty parameter. Note that in comparison with Problem (10.25), the constraints have been accommodated into the objective function via an \( \ell_1 \) penalty term and have been replaced by a trust region constraint. Hence, the subproblem QP is always
feasible and bounded and has an optimum. To contend with the nondifferentiability of the objective function, the $\ell_1$ terms can be re-transferred into the constraints as in (10.11b). Similar to the PSLP algorithm, if $d_k$ solves this problem along with Lagrange multiplier estimates $(u_{k+1}, v_{k+1})$, and if $x_{k+1} = x_k + d_k$ is $\epsilon$-feasible and satisfies the KKT conditions within a given tolerance, or if the fractional improvement in the original objective function is not better than a given tolerance over some $c$ consecutive iterations, the algorithm can be terminated. Otherwise, the process is repeated iteratively. This type of procedure enjoys the asymptotic local convergence properties of SQP methods but also achieves global convergence owing to the $\ell_1$ penalty function and the trust region features. However, it is also prone to the Maratos effect, and corrective measures are necessary to avoid this phenomenon. We refer the reader to the Notes and References section for further discussion on this topic.

10.5 Gradient Projection Method of Rosen

As we learned in Chapter 8, the direction of steepest descent is that of the negative gradient. In the presence of constraints, however, moving along the steepest descent direction may lead to infeasible points. The gradient projection method of Rosen [1960] projects the negative gradient in a way that improves the objective function while maintaining feasibility.

First, consider the following definition of a projection matrix.

10.5.1 Definition

An $n \times n$ matrix $P$ is called a projection matrix if $P = P^t$ and $PP = P$.

10.5.2 Lemma

Let $P$ be an $n \times n$ matrix. Then the following statements are true:

1. If $P$ is a projection matrix, $P$ is positive semidefinite.
2. $P$ is a projection matrix if and only if $I - P$ is a projection matrix.
3. Let $P$ be a projection matrix and let $Q = I - P$. Then $L = \{Px : x \in \mathbb{R}^n\}$ and $L^\perp = \{Qx : x \in \mathbb{R}^n\}$ are orthogonal linear subspaces.

Furthermore, any point $x \in \mathbb{R}^n$ can be represented uniquely as $p + q$, where $p \in L$ and $q \in L^\perp$.

**Proof**

Let $P$ be a projection matrix, and let $x \in \mathbb{R}^n$ be arbitrary. Then $x^tPx = x^tPPx = x^tP^tPx = \|Px\|^2 \geq 0$, and hence $P$ is positive semidefinite. This proves Part 1.

By Definition 10.5.1, Part 2 is obvious. Clearly, $L$ and $L^\perp$ are linear subspaces. Note that $P^tQ = P(I - P) = P - PP = 0$, and hence, $L$ and $L^\perp$ are
indeed orthogonal. Now let \( x \) be an arbitrary point in \( R^n \). Then \( x = 1x = (P + Q)x = Px + Qx = p + q \), where \( p \in L \) and \( q \in L^\perp \). To show uniqueness, suppose that \( x \) can also be represented as \( x = p' + q' \), where \( p' \in L \) and \( q' \in L^\perp \). By subtraction it follows that \( p - p' = q' - q \). Since \( p - p' \in L \) and \( q' - q \in L^\perp \), and since the only point in the intersection of \( L \) and \( L^\perp \) is the zero vector, it follows that \( p - p' = q' - q = 0 \). Thus, the representation of \( x \) is unique, and the proof is complete.

**Problems Having Linear Constraints**

Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad Ax \leq b \\
& \quad Qx = q,
\end{align*}
\]

where \( A \) is an \( m \times n \) matrix, \( Q \) is an \( \ell \geq n \) matrix, \( b \) is an \( m \)-vector, \( q \) is an \( \ell \)-vector, and \( f: R^n \to R \) is a different function. Given a feasible point \( x \), the direction of steepest descent is \(-\nabla f(x)\). However, moving along \(-\nabla f(x)\) may destroy feasibility. To maintain feasibility, \(-\nabla f(x)\) is projected so that we move along \( d = -PVf(x) \), where \( P \) is a suitable projection matrix. Lemma 10.5.3 gives the form of a suitable projection matrix \( P \) and shows that \(-PVf(x)\) is indeed an improving feasible direction, provided that \(-PVf(x) \neq 0\).

**10.5.3 Lemma**

Consider the problem to minimize \( f(x) \) subject to \( Ax \leq b \) and \( Qx = q \). Let \( x \) be a feasible point such that \( A_1x = b_1 \) and \( A_2x < b_2 \), where \( A' = (A_1', A_2') \) and \( b' = (b_1', b_2') \). Furthermore, suppose that \( f \) is differentiable at \( x \). If \( P \) is a projection matrix such that \( PVf(x) \neq 0 \), then \( d = -PVf(x) \) is an improving direction of \( f \) at \( x \). Furthermore, if \( M' = (A_1', Q') \) has full rank, and if \( P \) is of the form \( P = I - M'(MM')^{-1}M \), then \( d \) is an improving feasible direction.

**Proof**

Note that

\[
\nabla f(x)'d = -\nabla f(x)' PVf(x) = -\nabla f(x)' P' PVf(x) = -\|PVf(x)\|^2 < 0.
\]
By Lemma 10.1.2, $d = -PVf(x)$ is an improving direction. Furthermore, if $P = I - M'(MM')^{-1}M$, then $Md = -MPVf(x) = 0$; that is, $A_1d = 0$ and $Qd = 0$. By Lemma 10.1.2, $d$ is a feasible direction, and the proof is complete.

**Geometric Interpretation of Projecting the Gradient**

Note that the matrix $P$ of Lemma 10.5.3 is indeed a projection matrix satisfying $P = P'$ and $PP = P$. Furthermore, $MP = 0$; that is, $A_1P = 0$ and $Qp = 0$. In other words, the matrix $P$ projects each row of $A_1$ and each row of $Q$ into the zero vector. But since the rows of $A_1$ and $Q$ are the gradients of the binding constraints, $P$ is the matrix that projects the gradients of the binding constraints into the zero vector. Consequently, in particular, $PVf(x)$ is the projection of $Vf(x)$ onto the nullspace of the binding constraints.

Figure 10.15 illustrates the process of projecting the gradient for a problem having inequality constraints. At the point $x$, there is only one binding constraint with gradient $A_1$. Note that the matrix $P$ projects any vector onto the nullspace of $A_1$ and that $d = -PVf(x)$ is an improving feasible direction.

**Resolution of the Case $PVf(x) = 0$**

We have seen that if $PVf(x) \neq 0$, then $d = -PVf(x)$ is an improving feasible direction. Now, suppose that $PVf(x) = 0$. Then

$$0 = PVf(x) = [I - M'(MM')^{-1}M]Vf(x) = Vf(x) + M'w = Vf(x) + A_1u + Q'v,$$

Figure 10.15 Projecting the gradient.
where \( w = -(MM')^{-1} MVf(x) \) and \( w' = (u', v') \). If \( u \geq 0 \), then the point \( x \) satisfies the KKT conditions and we may stop. If \( u \not\geq 0 \), then as Theorem 10.5.4 shows, a new projection matrix \( \hat{P} \) can be identified such that \( d = -\hat{P} Vf(x) \) is indeed an improving feasible direction.

**10.5.4 Theorem**

Consider the problem to minimize \( f(x) \) subject to \( Ax \leq b \) and \( Qx = q \). Let \( x \) be a feasible solution, and suppose that \( A_1x = b_1 \) and \( A_2x < b_2 \), where \( A' = (A_1', A_2') \) and \( b' = (b_1', b_2') \). Suppose that \( M' = (A_1', Q') \) has full rank, and let \( P = I - M'(MM')^{-1}M \). Furthermore, suppose that \( PVf(x) = 0 \), and let \( w = -(MM')^{-1} MVf(x) \) and \( (u', v') = w' \). If \( u \geq 0 \), then \( x \) is a KKT point. If \( u \not\geq 0 \), let \( u_i \) be a negative component of \( u \), and let \( M' = (A_1', Q') \), where \( A_1 \) is obtained from \( A_1 \) by deleting the row of \( A_1 \) corresponding to \( u_j \). Now let \( \hat{P} = I - \hat{M}'(\hat{M}\hat{M}')^{-1}\hat{M} \), and let \( d = -\hat{P} Vf(x) \). Then \( d \) is an improving feasible direction.

**Proof**

By the definition of \( P \) and since \( PVf(x) = 0 \), we get

\[
0 = PVf(x) = [I - M'(MM')^{-1}M]Vf(x)
= Vf(x) + M'w = Vf(x) + A_1'u + Q'v.
\]

(10.33)

In view of (10.33), if \( u \geq 0 \), then \( x \) is a KKT point.

Now suppose that \( u \not\geq 0 \), and let \( u_j \) be a negative component of \( u \). Define \( \hat{P} \) as in the statement of the theorem. We first show that \( \hat{P} Vf(x) \not= 0 \). By contradiction, suppose that \( \hat{P} Vf(x) = 0 \). By the definition of \( \hat{P} \) and letting \( \hat{w} = -(\hat{M}\hat{M}')^{-1}\hat{M} Vf(x) \), we get

\[
0 = \hat{P} Vf(x) = [I - \hat{M}'(\hat{M}\hat{M}')^{-1}\hat{M}]Vf(x) = Vf(x) + \hat{M}'\hat{w}.
\]

(10.34)

Note that \( A_1'u + Q'v \) could be written as \( \hat{M}'\hat{w} + u_j r_j' \), where \( r_j \) is the \( j \)th row of \( A_1 \). Thus, from (10.33), we get

\[
0 = Vf(x) + \hat{M}'\hat{w} + u_j r_j'.
\]

(10.35)
Subtracting (10.35) from (10.34), it follows that $0 = \hat{M}'(\hat{w} - \bar{w}) - u_j r'_j$. This, together with the fact that $u_j \neq 0$, violates the assumption that $M$ has full rank. Therefore, $\hat{P}Vf(x) \neq 0$. Consequently, by Lemma 10.5.3, $d$ is an improving direction.

Now we show that $d$ is a feasible direction. Note that $\hat{M}\hat{P} = 0$, so that

$$
\begin{pmatrix}
\hat{A}_1 \\
\hat{Q}
\end{pmatrix}d = \hat{M}d = -\hat{M}\hat{P}Vf(x) = 0.
$$

(10.36)

By Lemma 10.5.3, $d$ is feasible direction if $A_1d \leq 0$ and $Qd = 0$. In view of (10.36), to show that $d$ is a feasible direction, it suffices to demonstrate that $r_jd \leq 0$. Premultiplying (10.35) by $r_j\hat{P}$, and noting that $PM' = 0$, it follows that

$$
0 = r_j\hat{P}Vf(x) + r_j\hat{P}(\hat{M}'\hat{w} + u_j r'_j) = -r_jd + u_j r_j r_j \hat{P}r'_j.
$$

By Lemma 10.5.2, $\hat{P}$ is positive semidefinite, so that $r_j\hat{P}r'_j \geq 0$. Since $u_j < 0$, the above equation implies that $r_jd \leq 0$. This completes the proof.

**Summary of the Gradient Projection Method of Rosen (Linear Constraints)**

We summarize below Rosen's gradient projection method for solving a problem of the form to minimize $f(x)$ subject to $Ax \leq b$ and $Qx = q$. We assume that for any feasible solution, the set of binding constraints are linearly independent. Otherwise, when the active constraints are dependent, $MM'$ is singular and the main algorithmic step is not defined. Moreover, in such a case, the Lagrange multipliers are nonunique, and an arbitrary choice of dropping a constraint can cause the algorithm to be stuck at a current, non-KKT solution.

**Initialization Step**  Choose a point $x_1$ satisfying $Ax_1 \leq b$ and $Qx_1 = q$. Suppose that $A'$ and $b'$ are decomposed into $(A'_1, A'_2)$ and $(b'_1, b'_2)$ such that $A_1x_1 = b_1$ and $A_2x_1 < b_2$. Let $k = 1$ and go to the Main Step.

**Main Step**

1. Let $M' = (A'_1, Q')$. If $M$ is vacuous, stop if $Vf(x_k) = 0$; else, let $d_k = -Vf(x_k)$, and proceed to Step 2. Otherwise, let $P = I - M'(MM')^{-1}M$ and set $d_k = -VPf(x_k)$. If $d_k \neq 0$, go to Step 2. If $d_k = 0$, compute $w = -(MM')^{-1}MVf(x_k)$ and let $w' = (u', v')$. If $u \geq 0$, stop; $x_k$ is a KKT point, with $w$ yielding the associated
Lagrange multipliers. If \( u \geq 0 \), choose a negative component of \( u \), say \( u_j \). Update \( A_1 \) by deleting the row corresponding to \( u_j \) and repeat Step 1.

2. Let \( \lambda_k \) be an optimal solution to the following line search problem:

\[
\begin{align*}
\text{Minimize} & \quad f(x_k + \lambda d_k) \\
\text{subject to} & \quad 0 \leq \lambda \leq \lambda_{\text{max}},
\end{align*}
\]

where \( \lambda_{\text{max}} \) is given by (10.1). Let \( x_{k+1} = x_k + \lambda_k d_k \), and suppose that \( A' \) and \( b' \) are decomposed into \( (A'_1, A'_2) \) and \( (b'_1, b'_2) \) such that \( A_1 x_{k+1} = b_1 \) and \( A_2 x_{k+1} < b_2 \). Replace \( k \) by \( k + 1 \) and go to Step 1.

10.5.5 Example

Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad 2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2 \\
\text{subject to} & \quad x_1 + x_2 \leq 2 \\
& \quad x_1 + 5x_2 \leq 5 \\
& \quad -x_1 \leq 0 \\
& \quad -x_2 \leq 0.
\end{align*}
\]

Note that \( \nabla f(x) = (4x_1 - 2x_2 - 4, 4x_2 - 2x_1 - 6)' \). We solve this problem using the gradient projection method of Rosen, starting from the point \((0, 0)\). At each iteration, we first find the direction to move by Step 1 of the algorithm and then perform a line search along this direction.

**Iteration 1:**

**Search Direction** At \( x_1 = (0, 0)' \), we have \( \nabla f(x_1) = (-4, -6)' \). Furthermore, only the nonnegativity constraints are binding at \( x_1 \), so that

\[
\begin{align*}
A_1 &= \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, & A_2 &= \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}.
\end{align*}
\]

We then have

\[
P = I - A_1(A_1'A_1)^{-1}A_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}
\]

and \( d_1 = -P\nabla f(x_1) = (0, 0)' \). Noting that we do not have equality constraints in this problem, we compute

\[
w = u = (A_1'A_1)^{-1}A_1\nabla f(x_1) = (-4, -6)'.
\]
Choosing $u_4 = -6$, we delete the corresponding gradient of the fourth constraint from $A_4$. The matrix $A_4$ is modified to give $\hat{A}_1 = (-1, 0)$. The modified projection matrix is then

$$\hat{P} = I - \hat{A}_1^T (\hat{A}_1 \hat{A}_1^T)^{-1} \hat{A}_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

and the direction $d_1$ to move is given by

$$d_1 = -\hat{P} \nabla f(x_1) = -\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -4 \\ -6 \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \end{bmatrix}.$$ 

**Line Search** Any point $x_2$ in the direction $d_1$ starting from the point $x_1$ can be written as $x_2 = x_1 + \lambda d_1 = (0, 6\lambda)'$, and the corresponding objective function value is $f(x_2) = 72\lambda^2 - 36\lambda$. The maximum value of $\lambda$ for which $x_k + \lambda d_1$ is feasible is obtained from (10.1) as

$$\lambda_{\text{max}} = \min \left\{ \frac{2}{6}, \frac{5}{30} \right\} = \frac{1}{6}.$$

Therefore, $\lambda_4$ is the optimal solution to the following problem:

Minimize $72\lambda^2 - 36\lambda$

subject to $0 \leq \lambda \leq \frac{1}{6}$.

The optimal solution is $\lambda_4 = 1/6$, so that $x_2 = x_1 + \lambda_4 d_1 = (0, 1)'$.

**Iteration 2:**

**Search Direction** At the point $x_2 = (0, 1)'$, we have $\nabla f(x_2) = (-6, -2)'$. Furthermore, at this point, constraints 2 and 3 are binding, so that we get

$$A_1 = \begin{bmatrix} 1 & 5 \\ -1 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}.$$ 

We then have

$$P = I - A_1^T (A_1 A_1^T)^{-1} A_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and hence $-P \nabla f(x_2) = (0, 0)'$. Thus, we compute
\[ u = -(A_1A_1^t)^{-1}A_1\nabla f(x_2) = \left( \frac{2}{5}, -\frac{28}{5} \right)' \].

Since \( u_3 < 0 \), the row \((-1, 0)\) is deleted from \( A_1 \), which gives the modified matrix \( \hat{A}_1 = [1, 5] \). The projection matrix and the corresponding direction vector are given by

\[ \hat{P} = I - \hat{A}_1^t(\hat{A}_1\hat{A}_1^t)^{-1}\hat{A}_1 = \begin{bmatrix} 25/26 & -5/26 \\ 5/26 & 1/26 \end{bmatrix} \]

\[ d_2 = -\hat{P}\nabla f(x_2) = \begin{bmatrix} 70/13 \\ 14/13 \end{bmatrix} \].

Since the norm of \( d_2 \) is not important, \((70/13, -14/13)'\) is equivalent to \((5, -1)'\). We therefore let \( d_2 = (5, -1)'\).

**Line Search** We are interested in points of the form \( x_2 + \lambda d_2 = (5\lambda, 1 - \lambda)' \) and \( f(x_2 + \lambda d_2) = 62\lambda^2 - 28\lambda - 4 \). The maximum value of \( \lambda \) for which \( x_2 + \lambda d_2 \) is feasible is obtained from (10.1) as

\[ \lambda_{\text{max}} = \min \left\{ \frac{1}{4}, 1 \right\} = \frac{1}{4}. \]

Therefore, \( \lambda_2 \) is the solution to the following problem:

\[
\text{Minimize } 62\lambda^2 - 28\lambda - 4 \\
\text{subject to } 0 \leq \lambda \leq \frac{1}{4}.
\]

The optimal solution is \( \lambda = 7/31 \), so that \( x_3 = x_2 + \lambda_2 d_2 = (35/31, 24/31)' \).

**Iteration 3:**

**Search Direction** At the point \( x_3 = (35/31, 24/31)' \), we have \( \nabla f(x_3) = (-32/31, -160/31)' \). Furthermore, the second constraint is binding, so that
Furthermore, we get

\[ P = I - A_1^T(A_1A_1^T)^{-1}A_1 = \frac{1}{26} \begin{bmatrix} 25 & -5 \\ -5 & 1 \end{bmatrix} \]

and the direction \( d_3 = -P \nabla f(x_3) = (0, 0)' \). Thus, we compute

\[ u = -(A_1A_1^T)^{-1}A_1 \nabla f(x_3) = \frac{32}{31} \geq 0. \]

Hence, the point \( x_3 \) is a KKT point. Note that the gradient of the binding constraint points in the direction opposite to \( \nabla f(x_3) \). In particular, \( \nabla f(x_3) + u_2 \nabla g_2(x_3) = 0 \) for \( u_2 = 32/31 \), thus verifying that \( x_3 \) is a KKT point. In this particular example, since \( f \) is strictly convex, then, by Theorem 4.3.8, the point \( x_3 \) is indeed the global optimal solution to the problem.

Table 10.4 summarizes the computations for solving the above problem. The progress of the algorithm is shown in Figure 10.16.

**Nonlinear Constraints**

So far we have discussed the gradient projection method for the case of linear constraints. In this case the projection of the gradient of the objective function onto the nullspace of the gradients of the binding constraints, or a subset of the

![Figure 10.16 Gradient projection method of Rosen](image)
Table 10.4 Summary of Computations for the Gradient Project Method of Rosen

<table>
<thead>
<tr>
<th>Iteration $k$</th>
<th>$x_k$</th>
<th>$f(x_k)$</th>
<th>$\nabla f(x_k)$</th>
<th>$I$</th>
<th>$A_1$</th>
<th>$P$</th>
<th>$d_k$</th>
<th>$u$</th>
<th>$\lambda_{max}$</th>
<th>$\lambda_k$</th>
<th>$x_{k+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0, 0)</td>
<td>0</td>
<td>(-4, -6)</td>
<td>(3, 4)</td>
<td>[-1 0]</td>
<td>[0 0]</td>
<td>[0 0]</td>
<td>(-4, -6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[0 1]</td>
<td>[0 0]</td>
<td>(0, 0)</td>
<td></td>
<td>$\frac{1}{6}$</td>
<td>$\frac{1}{6}$</td>
<td>(0, 1)</td>
</tr>
<tr>
<td>2</td>
<td>(0, 1)</td>
<td>-4.00</td>
<td>(-6, -2)</td>
<td>(2, 3)</td>
<td>[5 0]</td>
<td>[0 0]</td>
<td>(0, 0)</td>
<td>$\left(\frac{2}{5}, -\frac{28}{5}\right)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[1 0]</td>
<td>[0 0]</td>
<td>(0, 0)</td>
<td></td>
<td>$\frac{1}{4}$</td>
<td>$\frac{7}{31}$</td>
<td>(35, 24)</td>
</tr>
<tr>
<td>3</td>
<td>$\left(\frac{35}{31}, \frac{24}{31}\right)$</td>
<td>-7.16</td>
<td>$\left(-\frac{32}{31}, -\frac{160}{31}\right)$</td>
<td>(2, 3)</td>
<td>[5 0]</td>
<td>[0 0]</td>
<td>(0, 0)</td>
<td>$\left(\frac{32}{31}\right)$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
binding constraints, led to an improving feasible direction or to the conclusion that a KKT point was at hand. The same strategy can be used in the presence of nonlinear constraints. The projected gradient will usually not lead to feasible points, since it is only tangential to the feasible region, as illustrated in Figure 10.17. Therefore, a movement along the projected gradient must be coupled with a correction move to the feasible region.

To be more specific, consider the following problem:

\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0 \quad \text{for} \quad i = 1, \ldots, m \\
& \quad h_i(x) = 0 \quad \text{for} \quad i = 1, \ldots, \ell.
\end{align*}

Let $x_k$ be a feasible solution, and let $I = \{ i : g_i(x_k) = 0 \}$. Let $M$ be the matrix whose rows are $\nabla g_i(x_k)'$ for $i \in I$ and $\nabla h_i(x_k)'$ for $i = 1, \ldots, \ell$, and let $P = I - M'(MM')^{-1}M$. Note that $P$ projects any vector onto the nullspace of the gradients of the equality constraints and the binding inequality constraints. Let $d_k = -P\nabla f(x_k)$. If $d_k \neq 0$, then we minimize $f$ starting from $x_k$ in the direction $d_k$ and make a correction move to the feasible region. If, on the other hand, $d_k = 0$, then we calculate $(u', v') = -\nabla f(x_k)'M'(MM')^{-1}$. If $u \geq 0$, then we stop with a KKT point $x_k$. Otherwise, we delete the row of $M$ corresponding to some $u_i < 0$ and repeat the process.

**Convergence Analysis of the Gradient Projection Method**

Let us first examine the question whether the direction-finding map is closed or not. Note that the direction generated could change abruptly when a new restriction becomes active, or when the projected gradient is the zero vector, necessitating the computation of a new projection matrix. Hence, as shown below, this causes the direction-finding map to be *not* closed.

**10.5.6 Example**

Consider the following problem:
Minimize \( x_1 - 2x_2 \) subject to \( x_1 + 2x_2 \leq 6 \), \( x_1, x_2 \geq 0 \).

We now illustrate that the direction-finding map of the gradient projection method is not closed in general. Consider the sequence \( \{x_k\} \), where \( x_k = (2 - 1/k, 2) \). Note that \( \{x_k\} \) converges to the point \( \hat{x} = (2, 2) \). For each \( k \), \( x_k \) is feasible and the set of binding constraints is empty. Thus, the projection matrix is equal to the identity so that \( d_k = -\nabla f(x_k) = (-1, 2) \). Note, however, that the first constraint is binding at \( \hat{x} \). Here, the projection matrix is

\[
P = \begin{pmatrix}
4 & 2 \\
5 & 5 \\
2 & 1 \\
5 & 5
\end{pmatrix}
\]

and hence,

\[
d = -P\nabla f(\hat{x}) = \begin{pmatrix}
8 \\
5 \\
4 \\
5
\end{pmatrix}.
\]

Thus, \( \{d_k\} \) does not converge to \( d \), and the direction-finding map is not closed at \( \hat{x} \). This is illustrated in Figure 10.18. Not only is the direction-finding map not closed, but also the line search map that restricts the maximum step length via some feasible set is not closed in general, as seen in Example 10.2.2. Hence, Theorem 7.2.3 cannot be used to prove the convergence of this method. Nonetheless, one can prove that this algorithm converges under the following modification.

Figure 10.18 The direction-finding map is not closed.
Direction-Finding Routine for a Convergent Variant of the Gradient Projection Method

Consider the following revision of Step 1 of the Main Step of the gradient projection method summarized above for the case of linear constraints.

1. Let $M' = (A', Q')$. If $M$ is vacuous, stop if $\nabla f(x_k) = 0$, or else let $d_k = -\nabla f(x_k)$ and proceed to Step 2. Otherwise, let $P = I - M'(MM')^{-1}M$ and set $d_k^I = -P\nabla f(x_k)$. Also, compute $w = -(MM')^{-1}M\nabla f(x_k)$ and let $w' = (u', v')$. If $u \geq 0$, then stop if $d_k^I = 0$; otherwise, put $d_k = d_k^I \neq 0$ and proceed to Step 2. On the other hand, if $u \leq 0$, let $u_h = \min_j \{u_j\} < 0$, let $M' = (\hat{A}_1', Q')$, where $\hat{A}_1$ is obtained from $A_1$ by deleting the row of $A_1$ corresponding to $u_h$, construct the projection matrix $\hat{P} = I - \hat{M}'(\hat{M}\hat{M}')^{-1}\hat{M}$, and define $d_k^\Pi = -\hat{P}\nabla f(x_k)$. Now, based on some constant scalar $c > 0$, let

$$d_k = \begin{cases} d_k^I & \text{if } \|d_k^I\| > |u_h|c \\ d_k^\Pi & \text{otherwise} \end{cases} \tag{10.37}$$

and proceed to Step 2.

Note that if either $M$ is vacuous, or if $d_k^I = 0$ above, the procedural steps are the same as before. Hence, suppose that $M$ is nonvacuous and that $d_k^I \neq 0$. Whereas in the previous case, we would have used $d_k = d_k^I$ at this point, we now compute $w$ and switch to using $d_k^\Pi$ instead, provided that it turns out that $u \geq 0$ and that $\|d_k^I\|$ is "too small" by the measure $\|d_k^I\| \leq |u_h|c$. In particular, if $c = 0$, then Step 1 is identical for the two procedures. The following result establishes that Step 1 indeed generates an improving feasible direction.

10.5.7 Theorem

Consider the foregoing modification of Step 1 of the gradient projection method. Then either the algorithm terminates with a KKT solution at this step, or else it generates an improving feasible direction.

Proof

By Theorem 10.5.4, if the process stops at this step, it does so with a KKT solution. Also, from the above discussion, the claim follows from Theorem...
10.5.4 when $M$ is vacuous, or if $d_k^I = 0$, or if $u \geq 0$, or if $\|d_k^I\| > |u_h|c$. Hence, suppose that $M$ is nonvacuous, $u \geq 0$, and that $d_k^I \neq 0$, but $\|d_k^I\| < |u_h|c$, so that, by (10.37), we use $d_k = d_k^II$.

To begin with, note that $d_k^I = -\hat{P}Vf(x_k) \neq 0$, or else, by (10.34), we would have $d_k^I = -PVf(x_k) = PM'\hat{w} = 0$, since $MP' = MP = 0$ because $MP = 0$. This contradicts $d_k^I \neq 0$. Hence, $\hat{P}Vf(x_k) \neq 0$; so by Lemma 10.5.3, $d_k^I$ is an improving direction of $f$ at $x_k$.

Next, let us show that $d_k^I$ is a feasible direction. As in the proof of Theorem 10.5.4, noting (10.36), it suffices to demonstrate that $r_hd_k^I \leq 0$, where $r_h$ corresponds to the deleted row of $A_1$. As in (10.33) and (10.35), we have

$$PVf(x_k) = Vf(x_k) + M'w + rhPrf.$$ 

Premultiplying this by $r_h\hat{P}$ gives

$$r_h\hat{P}PVf(x_k) = -r_hd_k^I + r_h\hat{M}'\hat{w} + u_hr_hr_f.$$ 

Since $MP = 0$, we have $\hat{P}P = P$, so $r_h\hat{P}P = r_hP = 0$ as $MP = 0$. Also, since $\hat{P}M' = 0$, (10.38) yields $r_hd_k^I = u_hr_h\hat{P}r_f \leq 0$, since $u_h < 0$, and $\hat{P}$ is positive semidefinite by Lemma 10.5.2. This completes the proof.

Hence, by Theorem 10.5.7, the various steps of the algorithm are well defined. Although the direction finding and the line search maps are still not closed (Example 10.5.6 continues to apply), Du and Zhang [1989] demonstrate that convergence obtains with the foregoing modification by showing that if the iterates get too close within a defined $\varepsilon$-neighborhood of a non-KKT solution, every subsequent step changes the active constraint set by 1 until the iterates are forced out of this neighborhood. This is shown to occur in a manner that precludes a non-KKT point from becoming a cluster point of the sequence generated. We refer the reader to their paper for further details.

10.6 Reduced Gradient Method of Wolfe and Generalized Reduced Gradient Method

In this section we describe another procedure for generating improving feasible directions. The method depends upon reducing the dimensionality of the problem by representing all the variables in terms of an independent subset of the variables. The reduced gradient method was developed by Wolfe [1963] to solve a nonlinear programming problem having linear constraints. The method was later generalized by Abadie and Carpentier [1969] to handle nonlinear constraints. Consider the following problem.
Minimize $f(x)$
subject to $Ax = b$
$x \geq 0,$

where $A$ is an $m \times n$ matrix of rank $m$, $b$ is an $m$-vector, and $f$ is a continuously differentiable function on $R^m$. The following *nondegeneracy assumption* is made. Any $m$ columns of $A$ are linearly independent, and every extreme point of the feasible region has $m$ strictly positive variables. With this assumption, every feasible solution has at least $m$ positive components and, at most, $n - m$ zero components.

Now let $x$ be a feasible solution. By the nondegeneracy assumption, note that $A$ can be decomposed into $[B, N]$ and $x'$ into $[x_B', x_N']$, where $B$ is an $m \times m$ invertible matrix and $x_B > 0$. Here $x_B$ is called the *basic vector*, and each of its components is strictly positive. The components of the *nonbasic vector* $x_N$ may be either positive or zero. Let $\nabla f(x)' = [\nabla_B f(x)', \nabla_N f(x)']$, where $\nabla_B f(x)$ is the gradient of $f$ with respect to the basic vector $x_B$, and $\nabla_N f(x)$ is the gradient of $f$ with respect to the nonbasic vector $x_N$. Recall that a direction $d$ is an improving feasible direction of $f$ at $x$ if $\nabla f(x)'d < 0$, and if $Ad = 0$ with $d_j \geq 0$ if $x_j = 0$. We now specify a direction vector $d$ satisfying these properties.

First, $d'$ is decomposed into $[d_B', d_N']$. Note that $0 = Ad = Bd_B + Nd_N$ holds true automatically if for any $d_N$, we let $d_B = -B^{-1}Nd_N$. Let $r' = (r_B', r_N') = \nabla f(x)' - \nabla_B f(x)'B^{-1}A = [0, \nabla_N f(x)' - \nabla_B f(x)'B^{-1}N]$ be the *reduced gradient*, and let us examine the term $\nabla f(x)'d$:

$$\nabla f(x)'d = \nabla_B f(x)'d_B + \nabla_N f(x)'d_N = [\nabla_N f(x)' - \nabla_B f(x)'B^{-1}N]d_N = r_N'd_N.$$ 

We must choose $d_N$ in such a way that $r_N'd_N < 0$ and that $d_j \geq 0$ if $x_j = 0$.

The following rule is adopted. For each nonbasic $j$, let $d_j = -r_j$ if $r_j \leq 0$, and let $d_j = -x_jr_j$ if $r_j > 0$. This ensures that $d_j \geq 0$ if $x_j = 0$, and prevents unduly small step sizes when $x_j > 0$, but small, while $r_j > 0$. This also helps make the direction-finding map closed, thereby enabling convergence.

Furthermore, $\nabla f(x)'d \leq 0$, where strict inequality holds if $d_N \neq 0$.

To summarize, we have described a procedure for constructing an improving feasible direction. This fact, as well as the fact that $d = 0$ if and only if $x$ is KKT point, is proved in Theorem 10.6.1.
10.6.1 Theorem

Consider the problem to minimize $f(x)$ subject to $Ax = b$, $x \geq 0$, where $A$ is an $m \times n$ matrix and $b$ is an $m$-vector. Let $x$ be a feasible solution such that $x^t = (x_B^t, x_N^t)$ and $x_B > 0$, where $A$ is decomposed into $[B, N]$ and $B$ is an $m \times m$ invertible matrix. Suppose that $f$ is differentiable at $x$, and let $r^t = \nabla f(x)^t - \nabla_B f(x)^t B^{-1} A$. Let $d^t = (d_B^t, d_N^t)$ be the direction formed as follows. For each nonbasic component $j$, let $d_j = -r_j$ if $r_j \leq 0$ and $d_j = -x_j r_j$ if $r_j > 0$, and let $d_B = -B^{-1} Nd_N$. If $d \neq 0$, then $d$ is an improving feasible direction. Furthermore, $d = 0$ if and only if $x$ is a KKT point.

Proof

First, note that $d$ is a feasible direction if and only if $Ad = 0$, and $d_j \geq 0$ if $x_j = 0$ for $j = 1, ..., n$. By the definition of $d_B$, $Ad = Bd_B + Nd_N = B(-B^{-1} Nd_N) + Nd_N = 0$. If $x_j$ is basic, then $x_j > 0$ by assumption. If $x_j$ is not basic, then $d_j$ could be negative only if $x_j > 0$. Thus, $d_j \geq 0$ if $x_j = 0$, and hence, $d$ is a feasible direction. Furthermore,

$$\nabla f(x)^t d = \nabla_B f(x)^t d_B + \nabla_N f(x)^t d_N$$

$$= [\nabla_N f(x)^t - \nabla_B f(x)^t B^{-1} N]d_N = \sum_{j \in I} r_j d_j,$$

where $I$ is the index set of basic variables. Noting the definition of $d_j$, it is obvious that either $d = 0$ or $\nabla f(x)^t d < 0$. In the latter case, by Lemma 10.1.2, $d$ is indeed an improving feasible direction.

Note that $x$ is a KKT point if and only if there exist vectors $u^t = (u_B^t, u_N^t) \geq (0, 0)$ and $v$ such that

$$[\nabla_B f(x)^t, \nabla_N f(x)^t] + v^t (B, N) - (u_B^t, u_N^t) = (0, 0)$$

$$u_B^t x_B = 0, u_N^t x_N = 0.$$  

(10.39)

Since $x_B > 0$ and $u_B \geq 0$, then $u_B^t x_B = 0$ if and only if $u_B = 0$. From the first equation in (10.39), it follows that $v^t = -\nabla_B f(x)^t B^{-1}$. Substituting in the second equation in (10.39), it follows that $u_N^t = \nabla_N f(x)^t - \nabla_B f(x)^t B^{-1} N$. In other words, $u_N = r_N$. Thus, the KKT conditions reduce to $r_N \geq 0$ and $r_N^t x_N = 0$. By the definition of $d$, however, note that $d = 0$ if and only if $r_N \geq 0$ and $r_N^t x_N = 0$. Therefore, $d = 0$ if and only if $x$ is a KKT point.
Thus, $x$ is a KKT point if and only if $d = 0$, and the proof is complete.

Summary of the Reduced Gradient Algorithm

We summarize below Wolfe's reduced gradient algorithm for solving a problem of the form to minimize $f(x)$ subject to $Ax = b$, $x \geq 0$. It is assumed that all $m$ columns of $A$ are linearly independent and that every extreme point of the feasible region has $m$ strictly positive components. As we show shortly, the algorithm converges to a KKT point, provided that the basic variables are chosen to be the $m$ most positive variables, where a tie is broken arbitrarily.

Initialization Step Choose a point $x_1$ satisfying $Ax_1 = b$, $x_1 \geq 0$. Let $k = 1$ and go to the Main Step.

Main Step

1. Let $d_k' = (d_k', d_N')$ where $d_N'$ and $d_B'$ are obtained as below from (10.43) and (10.44), respectively. If $d_k' = 0$, stop; $x_k$ is a KKT point.

$[The \ Lagrange \ multipliers \ associated \ with \ Ax = b \ and \ x \geq 0 \ are, \ respectively, \ \nabla_B f(x_k)'B^{-1} \ and \ r.]$ Otherwise, go to Step 2.

$$I_k = \text{index set of the } m \ \text{largest components of } x_k \quad (10.40)$$

$$B = \{a_j : j \in I_k\}, \quad N = \{a_j : j \not\in I_k\} \quad (10.41)$$

$$r' = \nabla f(x_k)' - \nabla_B f(x_k)'B^{-1}A \quad (10.42)$$

$$d_j' = \begin{cases} -r_j & \text{if } j \not\in I_k \text{ and } r_j \leq 0 \\ -x_jr_j & \text{if } j \not\in I_k \text{ and } r_j > 0 \end{cases} \quad (10.43)$$

$$d_B' = -B^{-1}Nd_N'. \quad (10.44)$$

2. Solve the following line search problem:

$$\begin{align*}
\text{Minimize} & \quad f(x_k + \lambda d_k) \\
\text{subject to} & \quad 0 \leq \lambda \leq \lambda_{\text{max}},
\end{align*}$$

where

$$\lambda_{\text{max}} = \min_{1 \leq j \leq n} \left\{ \frac{-x_j}{d_{jk}} : d_{jk} < 0 \right\} \quad \text{if } d_k \not\geq 0$$

$$\infty \quad \text{if } d_k \geq 0 \quad (10.45)$$
and \( x_{jk}, d_{jk} \) are the \( j \)th components of \( x_k \) and \( d_k \), respectively. Let \( \lambda_k \) be an optimal solution, and let \( x_{k+1} = x_k + \lambda_k d_k \). Replace \( k \) by \( k + 1 \) and go to Step 1.

### 10.6.2 Example

Consider the following problem:

\[
\text{Minimize } 2x_1^2 + 2x_2^2 - 2x_1x_2 - 4x_1 - 6x_2 \\
\text{subject to } x_1 + x_2 + x_3 = 2 \\
\quad x_1 + 5x_2 + x_4 = 5 \\
\quad x_1, x_2, x_3, x_4 \geq 0.
\]

We solve this problem using Wolfe’s reduced gradient method starting from the point \( x_1 = (0, 0, 2, 5)^t \). Note that

\[
\nabla f(x) = (4x_1 - 2x_2 - 4x_2 - 2x_1 - 6, 0, 0)^t.
\]

We shall exhibit the information needed at each iteration in tableau form similar to the simplex tableau of Section 2.7. However, since the gradient vector changes at each iteration, and since the nonbasic variables could be positive, we explicitly give the gradient vector and the complete solution at the top of each tableau. The reduced gradient vector \( \mathbf{r}_k \) is shown as the last row of each tableau.

**Iteration 1:**

**Search Direction** At the point \( x_1 = (0, 0, 2, 5)^t \), we have \( \nabla f(x_1) = (-4, -6, 0, 0, 0) \). By (10.40), we have \( I_1 = \{3, 4\} \), so that \( B = [a_3, a_4] \) and \( N = [a_1, a_2] \). From (10.42), the reduced gradient is given by

\[
\mathbf{r}' = (-4, -6, 0, 0) - (0, 0)
\begin{bmatrix}
1 & 1 & 1 & 0 \\
1 & 5 & 0 & 1 \\
\end{bmatrix}
= (-4, -6, 0, 0).
\]

Note that the computations for the reduced gradient are similar to the computations for the objective row coefficients in the simplex method of Section 2.7. Also, \( r_i = 0 \) for \( i \in I_1 \). The information at this point is summarized in the following tableau.

<table>
<thead>
<tr>
<th>( \mathbf{x} )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Solution } x_1 )</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>( \nabla f(x_1) )</td>
<td>(-4)</td>
<td>(-6)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \nabla_B f(x_1) = \begin{bmatrix} 0 \ 0 \end{bmatrix} )</td>
<td>( x_3 )</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( \begin{bmatrix} 0 \ 0 \end{bmatrix} )</td>
<td>( x_4 )</td>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>( \mathbf{r} )</td>
<td>(-4)</td>
<td>(-6)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
By (10.16) we have \( \mathbf{d}_N = (d_1, d_2)' = (4, 6)' \). We now compute \( \mathbf{d}_B \) using (10.44) to get

\[
\mathbf{d}_B = (d_3, d_4)' = -\mathbf{B}^{-1}\mathbf{N} \mathbf{d}_N = -\begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} 4 \\ 6 \end{bmatrix} = (-10, -34)'.
\]

Note that \( \mathbf{B}^{-1}\mathbf{N} \) is recorded under the variables corresponding to \( \mathbf{N} \) : namely, \( x_1 \) and \( x_2 \). The direction vector is, then, \( \mathbf{d}_1 = (4, 6, -10, -34)' \).

**Line Search** Starting from \( x_1 = (0, 0, 2, 5)' \), we now wish to minimize the objective function along the direction \( \mathbf{d}_1 = (4, 6, -10, -34)' \). The maximum value of \( \lambda \) such that \( x_1 + \lambda \mathbf{d}_1 \) is feasible is computed using (10.45), and we get

\[
\lambda_{\text{max}} = \min \left\{ \frac{2}{10}, \frac{5}{34} \right\} = \frac{5}{34}.
\]

The reader can verify that \( f(x_1 + \lambda \mathbf{d}_1) = 56\lambda^2 - 52\lambda \), so that \( \lambda_1 \) is the solution to the following problem:

Minimize \( 56\lambda^2 - 52\lambda \)

subject to \( 0 \leq \lambda \leq \frac{5}{34} \).

This yields \( \lambda_1 = 5/34 \), so that \( x_2 = x_1 + \lambda_1 \mathbf{d}_1 = (10/17, 15/17, 9/17, 0)' \).

**Iteration 2:**

**Search Direction** At \( x_2 = (10/17, 15/17, 9/17, 0)' \), from (10.40) we have \( I_2 = \{1, 2\} \), \( \mathbf{B} = [\mathbf{a}_1, \mathbf{a}_2] \), and \( \mathbf{N} = [\mathbf{a}_3, \mathbf{a}_4] \). We also have \( \nabla f(x_2) = (-58/17, -62/17, 0, 0)' \). The current information is recorded in the following tableau, where the rows of \( x_1 \) and \( x_2 \) were obtained by two pivot operations on the tableau for Iteration 1.

<table>
<thead>
<tr>
<th>( \nabla f(x_2) )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solution ( x_2 )</td>
<td>10/17</td>
<td>15/17</td>
<td>9/17</td>
<td>0</td>
</tr>
<tr>
<td>( \nabla_B f(x_2) )</td>
<td>\begin{bmatrix} -58/17 \ -62/17 \end{bmatrix}</td>
<td>\begin{bmatrix} 1 &amp; 0 &amp; 5/4 &amp; -1/4 \ 0 &amp; 1 &amp; -1/4 &amp; 1/4 \end{bmatrix}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( r )</td>
<td>0</td>
<td>0</td>
<td>57/17</td>
<td>1/17</td>
</tr>
</tbody>
</table>

We have from (10.42)
From (10.43), then, \(d_3 = -(9/17)(57/17) = -513/289\) and \(d_4 = 0\), so that \(d_N = (-513/289, 0)^T\). From (10.44), we get

\[
\begin{bmatrix}
\frac{5}{4} & -
\frac{1}{4} & -
\frac{513}{289} \\
-1 & 1 & 0 \\
-4 & 4 & 0
\end{bmatrix}
\begin{bmatrix}
d_1 \\
d_2 \\
d_3
\end{bmatrix}
= \begin{bmatrix}
\frac{2565}{1156} \\
\frac{-513}{1156} \\
\frac{-513}{1156}
\end{bmatrix}.
\]

The new search direction is therefore given by \(d_2 = (2565/1156, -513/1156, -513/289, 0)^T\).

**Line Search** Starting from \(x_2 = (10/17, 15/17, 9/17, 0)^T\), we wish to minimize the objective function along the direction \(d_2 = (2565/1156, -513/1156, -513/289, 0)^T\). The maximum value of \(\lambda\) such that \(x_2 + \lambda d_2\) is feasible is computed using (10.45), and we get

\[
\lambda_{\text{max}} = \min \left\{ \frac{-15/17}{-513/1156}, \frac{-9/17}{-513/289} \right\} = \frac{17}{57}.
\]

The reader can verify that \(f(x_2 + \lambda d_2) = 12.21 \lambda^2 - 5.95 \lambda - 6.436\), so that \(\lambda_2\) is obtained by solving the following problem:

Minimize \(12.21 \lambda^2 - 5.9 \lambda - 6.436\)

subject to \(0 \leq \lambda \leq \frac{17}{57}\).

This can be verified to yield \(\lambda_2 = 68/279\), so that \(x_3 = x_2 + \lambda_2 d_2 = (35/31, 24/31, 3/31, 0)^T\).

**Iteration 3:**

**Search Direction** Now \(I_3 = \{1, 2\}\), so that \(B = [a_1, a_2]\) and \(N = [a_3, a_4]\). Since \(I_3 = I_2\), the tableau at Iteration 2 can be retained. However, we now have \(\nabla f(x_3) = (-32/31, -160/31, 0, 0)^T\).
From (10.42) we get
\[ r' = \left[ \begin{array}{c} -32 \\ 31 \\ 160 \\ 31 \\ 0 \\ 0 \end{array} \right] \left[ \begin{array}{cccc} 1 & 0 & 5 & -1 \\ 4 & 4 & 4 & 4 \\ 1 & 1 & 1 & 1 \\ 4 & 4 & 4 & 4 \end{array} \right] = \left[ \begin{array}{c} 0 \\ 0 \\ 0 \\ 32 \\ 31 \end{array} \right]. \]

From (10.43), \( d_N = (d_3, d_4)' = (0, 0)' \); and from (10.44) we also get \( d_B = (d_1, d_2)' = (0, 0)' \). Hence, \( d = 0 \), and the solution \( x_3 \) is a KKT solution and therefore optimal for this problem. The optimal Lagrange multipliers associated with the equality constraints are \( V_B f(x_3)' B^{-1} = (0, -32/31)' \), and those associated with the nonnegativity constraints are \( (0, 0, 0, 1)' \). Table 10.5 gives a summary of the computations, and the progress of the algorithm is shown in Figure 10.19.

**Convergence of the Reduced Gradient Method**

Theorem 10.6.3 proves convergence of the reduced gradient method to a KKT point. This is done by a contradiction argument that establishes a sequence satisfying conditions 1 through 4 of Lemma 10.2.6.

**10.6.3 Theorem**

Let \( f: R^n \to R \) be continuously differentiable, and consider the problem to minimize \( f(x) \) subject to \( Ax = b, x \geq 0 \). Here \( A \) is an \( m \times n \) matrix and \( b \) is an \( m \)-vector such that all extreme points of the feasible region have \( m \) positive components, and any set of \( m \) columns of \( A \) are linearly independent. Suppose that the sequence \( \{x_k\} \) is generated by the reduced gradient algorithm. Then any accumulation point of \( \{x_k\} \) is a KKT point.
Figure 10.19 Illustration of the reduced gradient method of Wolfe.

Proof

Let \( \{x_k\}_{k \in \mathcal{K}} \) be a convergent subsequence with limit \( \hat{x} \). We need to show that \( \hat{x} \) is a KKT point. Suppose, by contradiction, that \( \hat{x} \) is not a KKT point. We shall exhibit a sequence \( \{(x_k, d_k)\}_{k \in \mathcal{K}} \), satisfying conditions 1 through 4 of Lemma 10.2.6, which is impossible.

Let \( \{d_k\}_{k \in \mathcal{K}} \) be the sequence of directions associated with \( \{x_k\}_{k \in \mathcal{K}} \). Note that \( d_k \) is defined by (10.40) through (10.44) at \( x_k \). Letting \( I_k \) be the set denoting the indices of the \( m \) largest components of \( x_k \) used to compute \( d_k \), there exists \( \mathcal{K}_k' \subseteq \mathcal{K} \) such that \( I_k = \tilde{I} \) for each \( k \in \mathcal{K}_k' \), where \( \tilde{I} \) is the set denoting the indices of the \( m \) largest components of \( \hat{x} \). Let \( \tilde{d} \) be the direction obtained from (10.40) through (10.44) at \( k \), and note, by Theorem 10.6.1, that \( \tilde{d} \neq 0 \) and \( \nabla f(\hat{x})^T \tilde{d} < 0 \). Since \( f \) is continuously differentiable, \( x_k \to \hat{x} \) and \( I_k = \tilde{I} \) for \( k \in \mathcal{K}_k' \), then, by (10.41)-(10.44), \( d_k \to \tilde{d} \) for \( k \in \mathcal{K}_k' \). To summarize, we have exhibited a sequence \( \{(x_k, d_k)\}_{k \in \mathcal{K}} \), satisfying conditions 1, 2, and 4 of Lemma 10.2.6. To complete the proof, we need to show that Part 3 also holds true.

From (10.45), recall that \( x_k + \lambda d_k \) is feasible for each \( \lambda \in [0, \delta_k] \), where \( \delta_k = \min\{\min\{-x_{pk}/d_{pk} : d_{pk} < 0\}, \infty\} > 0 \) for each \( k \in \mathcal{K}_k' \). Suppose that \( \inf \{\delta_k : k \in \mathcal{K}_k'\} = 0 \). Then there exists an index set \( \mathcal{K}_k'' \subseteq \mathcal{K}_k' \) such that \( \delta_k = -x_{pk}/d_{pk} \) converges to 0 for \( k \in \mathcal{K}_k'' \), where \( x_{pk} > 0, d_{pk} < 0 \), and \( p \) is an element of \( \{1, \ldots, n\} \). By (10.40)-(10.44), note that \( \{d_{pk}\}_{k \in \mathcal{K}_k''} \) is bounded; and
<table>
<thead>
<tr>
<th>Iteration $k$</th>
<th>$x_k$</th>
<th>$f(x_k)$</th>
<th>$r_k$</th>
<th>$d_k$</th>
<th>$\lambda_k$</th>
<th>$x_{k+1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$(0,0,2,5)$</td>
<td>0.0</td>
<td>$(-4,-6,0,0)$</td>
<td>$(4,6,-10,-34)$</td>
<td>5/34</td>
<td>$(10,15,9,0)_{17,17,17,0}$</td>
</tr>
<tr>
<td>2</td>
<td>$(10,15,9,0)_{17,17,17,0}$</td>
<td>-6.436</td>
<td>$(0,0,57,4)_{17,17}$</td>
<td>$(2565,-513,-513,0)_{1156,1156,289,0}$</td>
<td>68/279</td>
<td>$(35,24,3,0)_{31,31,31,0}$</td>
</tr>
<tr>
<td>3</td>
<td>$(35,24,3,0)_{31,31,31,0}$</td>
<td>-7.16</td>
<td>$(0,0,0,32)_{31}$</td>
<td>$(0,0,0,0)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
since \( \{ \delta_k \}_k \) converges to 0, \( \{ x_{pk} \}_k \) converges to 0. Thus, \( \hat{x}_p = 0 \), that is, \( p \notin \hat{I} \). But \( I_k = \hat{I} \) for \( k \in \mathbb{N}^* \), and hence, \( p \notin I_k \). Since \( d_{pk} < 0 \), from (10.43), \( d_{pk} = -x_{pk} r_{pk} \). It then follows that \( \delta_k = -x_{pk} d_{pk} = 1/r_{pk} \). This shows that \( r_{pk} \to \infty \), which is impossible, since \( r_{pk} \to r_p \neq \infty \). Thus, \( \inf \{ \delta_k : k \in \mathbb{N}^* \} = \delta > 0 \).

We have thus shown that there exists a \( \delta > 0 \) such that \( x_k + \lambda d_k \) is feasible for each \( \lambda \in [0, \delta] \) and for each \( k \in \mathbb{N}^* \). Therefore, Condition 3 of Lemma 10.2.6 holds true, and the proof is complete.

**Generalized Reduced Gradient Method**

We can extend the reduced gradient method to handle nonlinear constraints, similar to the gradient projection method. This extension is referred to as the generalized reduced gradient (GRG) method, and is sketched below briefly (see also Exercise 10.56 for the scheme proposed originally).

Consider a nonlinear programming problem of the form

\[
\text{Minimize } \{ f(x) : h(x) = 0, x \geq 0 \},
\]

where \( h(x) = 0 \) represents some \( m \) equality constraints, \( x \in \mathbb{R}^n \), and suitable variable transformations have been used to represent all variables as being nonnegative. Here, any inequality constraint can be assumed to have been written as an equality by introducing a nonnegative slack variable.

Now, given a feasible solution \( x_k \), consider a linearization of \( h(x) = 0 \) given by \( h(x_k) + \nabla h(x_k)(x - x_k) = 0 \), where \( \nabla h(x_k) \) is the \( m \times n \) Jacobian of \( h \) evaluated at \( x_k \). Noting that \( h(x_k) = 0 \), the set of linear constraints given by \( \nabla h(x_k) x = \nabla h(x_k) x_k \) is of the form \( Ax = b \), where \( x_k \geq 0 \) is a feasible solution.

Assuming that the Jacobian \( A = \nabla h(x_k) \) has full row rank, and partitioning it suitably into \( [B, N] \) and, accordingly partitioning \( x' = (x'_B, x'_N) \) (where, hopefully, \( x'_B > 0 \) in \( x_k \) ), we can compute the reduced gradient \( r \) via (10.42) and, hence, obtain the direction of motion \( d_k \) via (10.43) and (10.44). As before, we obtain \( d_k = 0 \) if and only if \( x_k \) is a KKT point, whence the procedure terminates. Otherwise, a line search is performed along \( d_k \).

Earlier versions of this method adopted the following strategy. First, a line search is performed by determining \( \lambda_{\max} \) via (10.45) and then finding \( \lambda_k \) as the solution to the line search problem to minimize \( f(x_k + \lambda d_k) \) subject to \( 0 \leq \lambda \leq \lambda_{\max} \). This gives \( x' = x_k + \lambda_k d_k \). Since \( h(x') = 0 \) is not necessarily satisfied, we need a correction step (see also Exercise 10.7). Toward this end, the Newton-Raphson method is then used to obtain \( x_{k+1} \) satisfying \( h(x_{k+1}) = 0 \), starting with the solution \( x' \) and keeping the components of \( x_N \) fixed at the values \( x'_N \). Hence, \( x_N \) remains at \( x'_N \geq 0 \) during this iterative process, but some
component(s) of $x_B$ may tend to become negative. At such a point, a switch is made by replacing a negative basic variable $x_r$ with a nonbasic variable $x_q$ that is preferably positive and that has a significantly nonzero element in the corresponding row $r$ of the column $B^{-1}a_q$. The Newton-Raphson process then continues as above with the revised basis (having now fixed $x_r$ at zero) and the revised linearized system, until a nonnegative solution $x_{k+1}$ satisfying $h(x_{k+1}) = 0$ is finally obtained.

More recent versions of the GRG method adopt a discrete sequence of positive step sizes and attempt to find a corresponding $x_{k+1}$ for each such step size sequentially using the foregoing Newton-Raphson scheme. Using the value $f(x_{k+1})$ at each such point, when a three-point pattern (TPP) of the quadratic interpolation method (see Section 8.3) is obtained, a quadratic fit is used to determine a new step size, for which the corresponding point $x_{k+1}'$ is again computed as above using the Newton-Raphson scheme. A feasible point having the smallest objective value thus found is used as the next iterate. This technique appears to yield a more reliable algorithm.

As the reader may have surmised, the iterative Newton-Raphson scheme complicates convergence arguments. Indeed, the existing convergence proofs use restrictive and difficult-to-verify assumptions. Nonetheless, this type of algorithm provides quite a robust and efficient scheme for solving nonlinear programming problems.

### 10.7 Convex-Simplex Method of Zangwill

The convex-simplex method is identical to the reduced gradient method of Section 10.6, except that only one nonbasic variable is modified while all other nonbasic variables are fixed at their current levels. Of course, the values of the basic variables are modified accordingly to maintain feasibility, so that the method behaves very much like the simplex method for linear programs. The name *convex-simplex method* was coined because the method was originally proposed by Zangwill [1967] for minimizing a convex function in the presence of linear constraints. Below we reconstruct this algorithm as a modification of the reduced gradient method for solving the following class of problems:

Minimize $f(x)$

subject to $Ax = b$

$x \geq 0$,

where $A$ is an $m \times n$ matrix of rank $m$ and $b$ is an $m$-vector.

### Summary of the Convex-Simplex Method

We again assume that any $m$ columns of $A$ are linearly independent and that every extreme point of the feasible region has $m$ strictly positive components. As we shall show shortly, the algorithm converges to a KKT point, provided that
the basic variables are chosen to be the $m$ most positive variables, where a tie is broken arbitrarily.

**Initialization Step** Choose a point $x_1$ such that $Ax_1 = b$ and $x_1 \geq 0$. Let $k = 1$ and go to the Main Step.

**Main Step**

1. Given $x_k$, identify $I_k$, $B$, $N$, and compute $r$ as follows:
   \[
   I_k = \text{index set of the } m \text{ largest components of } x_k \quad (10.46)
   \]
   \[
   B = \{a_j : j \in I_k\}, \quad N = \{a_j : j \notin I_k\} \quad (10.47)
   \]
   \[
   r' = \nabla f(x_k)' - \nabla_B f(x_k)'B^{-1}A. \quad (10.48)
   \]
   Consider (10.49)-(10.55). If $\alpha = \beta = 0$, stop; $x_k$ is a KKT point having Lagrange multipliers $\nabla_B f(x_k)'B^{-1}$ and $r$, respectively, associated with the constraints $Ax = b$ and $x \geq 0$. If $\alpha > \beta$, compute $d_N$ from (10.51) and (10.53). If $\alpha < \beta$, compute $d_N$ from (10.52) and (10.54). If $\alpha = \beta \neq 0$, compute $d_N$ either from (10.51) and (10.53) or else from (10.52) and (10.54). In all cases, determine $d_B$ from (10.55) and go to Step 2.
   \[
   \alpha = \max \{-r_j : r_j \leq 0\} \quad (10.49)
   \]
   \[
   \beta = \max \{xjr_j : r_j \geq 0\} \quad (10.50)
   \]
   \[
   \nu = \begin{cases} 
   \text{an index such that } \alpha = -r_\nu & \text{if } \alpha \geq \beta \text{ is invoked} \\
   \text{an index such that } \beta = x_\nu r_\nu & \text{if } \beta \geq \alpha \text{ is invoked} 
   \end{cases} \quad (10.51, 10.52)
   \]
   In case $\alpha \geq \beta$ is invoked: $d_j = \begin{cases} 
   0 & \text{if } j \notin I_k, j \neq \nu \\
   1 & \text{if } j \notin I_k, j = \nu 
   \end{cases} \quad (10.53)$
   In case $\beta \geq \alpha$ is invoked: $d_j = \begin{cases} 
   0 & \text{if } j \notin I_k, j \neq \nu \\
   -1 & \text{if } j \notin I_k, j = \nu 
   \end{cases} \quad (10.54)$
   \[
   d_B = -B^{-1}N d_N = -B^{-1}a_\nu d_\nu. \quad (10.55)
   \]

2. Consider the following line search problem:
   \[
   \text{Minimize } f(x_k + \lambda d_k) \\
   \text{subject to } 0 \leq \lambda \leq \lambda_{\max},
   \]
   where
\[ \lambda_{\text{max}} = \begin{cases} \min_{1 \leq j \leq n} \left\{ \frac{-x_{jk}}{d_{jk}} : d_{jk} < 0 \right\} & \text{if } d_k \geq 0 \\ \infty & \text{if } d_k \geq 0 \end{cases} \quad (10.56) \]

and \( x_{jk} \) and \( d_{jk} \) are the \( j \)th components of \( x_k \) and \( d_k \), respectively.

Let \( \lambda_k \) be an optimal solution, and let \( x_{k+1} = x_k + \lambda_k d_k \). Replace \( k \) by \( k+1 \) and go to Step 1.

Observe that \( \alpha = \beta = 0 \) if and only if \( \text{d}_N = 0 \) in the reduced gradient method, which by Theorem 10.6.1 happens if and only if \( x_k \) is a KKT point. Otherwise, \( \text{d} \neq 0 \) is an improving feasible direction as in the proof of Theorem 10.6.1.

**10.7.1 Example**

Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad 2x_1^2 + 2x_2^2 - 2x_1 x_2 - 4x_1 - 6x_2 \\
\text{subject to} & \quad x_1 + x_2 + x_3 = 2 \\
& \quad x_1 + 5x_2 + x_4 = 5 \\
& \quad x_1, x_2, x_3, x_4 \geq 0.
\end{align*}
\]

We solve the problem using Zangwill's convex simplex method starting from the point \( x_1 = (0, 0, 2, 5)' \). Note that

\[ \nabla f(x) = (4x_1 - 2x_2 - 4, 4x_2 - 2x_1 - 6, 0, 0)' . \]

As in the reduced gradient method, it is convenient to exhibit the information at each iteration in tableau form, giving the solution vector \( x_k \) and also \( \nabla f(x_k) \).

**Iteration 1:**

**Search Direction** At the point \( x_1 = (0, 0, 2, 5)' \), we have \( \nabla f(x_1) = (-4, -6, 0, 0)' \). From (10.46) we then have \( I_1 = \{3, 4\} \), so that \( B = [a_3, a_4] \) and \( N = [a_1, a_2] \).

The reduced gradient is computed using (10.48) as follows:

\[ r' = (-4, -6, 0, 0) - (0, 0) \begin{bmatrix} 1 & 1 & 1 \\ 1 & 5 & 0 \end{bmatrix} = (-4, -6, 0, 0). \]

The tableau at this stage is given below.
Now, from (10.49), \( a = \max\{-q, -r_2, -r_3, -r_4\} = -r_2 = 6 \). Also, from (10.50), \( \beta = \max\{x_3, r_3, x_4, r_4\} = 0 \); hence, from (10.51), \( v = 2 \). Note that \(-r_2 = 6\) implies that \( x_2 \) can be increased to yield a reduced objective function value. The search direction is given by (10.53) and (10.55). From (10.53) we have \( d_N^t = (d_1, d_2) = (0, 1) \); and from (10.55) we get \( d_B^t = (d_3, d_4) = (-1, 5) \). Note that \( d_B = -B^{-1}a_2 \) is the negative of the column of \( x_2 \) in the above tableau. Hence, \( d_1 = (0, 1, -1, -5)^t \).

**Line Search** Starting from the point \( x_1 = (0, 0, 2, 5)^t \), we wish to search along the direction \( d_1 = (0, 1, -1, -5)^t \). The maximum value of \( \lambda \) such that \( x_1 + \lambda d_1 \) is feasible is given by (10.56). In this case,

\[
\lambda_{\text{max}} = \min \left\{ \frac{2}{1}, \frac{5}{5} \right\} = 1.
\]

We also have \( f(x_1 + \lambda d_1) = 2\lambda^2 - 6\lambda \). Hence, we solve the following problem:

\[
\begin{align*}
\text{Minimize} \quad & 2\lambda^2 - 6\lambda \\
\text{subject to} \quad & 0 \leq \lambda \leq 1.
\end{align*}
\]

The optimal solution is \( \lambda = 1 \), so that \( x_2 = x_1 + \lambda d_1 = (0, 1, 1, 0)^t \).

**Iteration 2:**

**Search Direction** At the point \( x_2 = (0, 1, 1, 0)^t \), we have, by (10.46), \( I_2 = \{2, 3\} \), so that \( B = [a_2, a_3] \) and \( N = [a_1, a_4] \). The updated tableau obtained by one pivot operation is given below. Note that \( \nabla f(x_2) = (-6, -2, 0, 0)^t \); and, from (10.48), we get

\[
r^t = (-6, -2, 0, 0) - (0, -2) = \begin{bmatrix} 4 & 0 & -1 & 5 \\ 1 & 0 & 1 & 5 \\ 1 & 1 & 1 & 5 \end{bmatrix} = \begin{bmatrix} -28/5 \\ 0 \\ 2 \\ 5 \end{bmatrix}.
\]
Methods of Feasible Directions

<table>
<thead>
<tr>
<th>Solution $x_2$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Vf(x_2)$</td>
<td>-6</td>
<td>-2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

From (10.49) and (10.50), $\alpha = \max\{-\eta_1, -\eta_2, -\eta_3\} = -\eta_1 = 28/5$, and $\beta = \max\{x_2 r_2, x_3 r_3, x_4 r_4\} = 0$, so that $v = 1$. This means that $x_1$ can be increased.

From (10.53) and (10.55), we have $d_N^f = (d_1, d_4) = (1, 0)$, and $d_B^f = (d_3, d_2) = (-4/5, -1/5)$. Thus, $d_2 = (1, -1/5, -4/5, 0)^t$.

### Line Search

Starting from the point $x_2 = (0, 1, 1, 0)^t$, we wish to search along the direction $d_2 = (1, -1/5, -4/5, 0)^t$. The maximum value of $\lambda$ so that $x_2 + \lambda d_2$ is feasible is given by (10.56) as follows:

$$\lambda_{\text{max}} = \min \left\{ \frac{1}{1/5}, \frac{1}{4/5} \right\} = \frac{5}{4}.$$

We also have $f(x_2 + \lambda d_2) = 2.48 \lambda^2 - 5.6 \lambda - 4$. Hence, we solve the problem:

Minimize $2.48 \lambda^2 - 5.6 \lambda - 4$

subject to $0 \leq \lambda \leq \frac{5}{4}$.

The optimal solution is $\lambda_2 = 35/31$, so that $x_3 = x_2 + \lambda_2 d_2 = (35/31, 24/31, 3/31, 0)^t$.

### Iteration 3:

#### Search Direction

At the point $x_3 = (35/31, 24/31, 3/31, 0)^t$, from (10.46), we get $I_3 = \{1, 2\}$, so that $B = [a_1, a_2]$ and $N = [a_3, a_4]$. We also have $Vf(x_3) = (-32/31, -160/31, 0, 0)^t$, and from (10.48), we get
Chapter 10

\[ r' = \left( \begin{array}{c} \frac{32}{31} \\ \frac{160}{31} \\ 0 \\ 0 \end{array} \right) - \left( \begin{array}{c} \frac{32}{31} \\ \frac{160}{31} \\ 0 \\ 0 \end{array} \right) \begin{pmatrix} 1 & 0 & 5 & \frac{1}{4} \\ 0 & 1 & \frac{1}{4} & \frac{1}{4} \end{pmatrix} = \left( 0, 0, 0, \frac{32}{31} \right). \]

The information is given in the next tableau.

<table>
<thead>
<tr>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 35 )</td>
<td>( 24 )</td>
<td>( 3 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( 31 )</td>
<td>( 31 )</td>
<td>( 31 )</td>
<td></td>
</tr>
</tbody>
</table>

\[ \nabla f(x_3) = \begin{pmatrix} \frac{32}{31} \\ \frac{160}{31} \end{pmatrix}, \quad \nabla_B f(x_3) = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \]

In this case, \( \alpha = \text{maximum} \{ -\eta_1, -\eta_2, -\eta_3 \} = 0 \) and \( \beta = \text{maximum} \{ x_1 \eta_1, x_2 \eta_2, x_3 \eta_3, x_4 \eta_4 \} = 0 \). Hence, the point \( x_3 = (35/31, 24/31, 3/31, 0)' \) is a KKT solution and, therefore, is optimal for this problem. (The optimal Lagrange multipliers are obtained as in Example 10.6.2.) A summary of the computations is given in Table 10.6. The progress of the algorithm is depicted in Figure 10.20.

**Convergence of the Convex-Simplex Method**

The convergence of the convex-simplex method to a KKT point can be established by an argument similar to that in Theorem 10.6.3. For the sake of completeness, this argument is sketched below.

**Table 10.6 Summary of Computation for the Convex-Simplex Method of Zangwill**

<table>
<thead>
<tr>
<th>Iter. ( k )</th>
<th>( x_k )</th>
<th>( f(x_k) )</th>
<th>( r )</th>
<th>( d )</th>
<th>( \lambda_k )</th>
<th>( x_{k+1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0, 0, 2, 5)</td>
<td>0.0</td>
<td>((-4, -6, 0, 0))</td>
<td>(0, 1, -1, -5)</td>
<td>1</td>
<td>(0, 1, 1, 0)</td>
</tr>
<tr>
<td>2</td>
<td>(0, 1, 1, 0)</td>
<td>-4.0</td>
<td>(-\left( -\frac{28}{5}, 0, 0, \frac{2}{5} \right) )</td>
<td>( 1, -\frac{1}{5}, -\frac{4}{5}, 0 )</td>
<td>( \frac{35}{31} )</td>
<td>( \left( \frac{35}{31}, \frac{24}{31}, \frac{3}{31}, 0 \right) )</td>
</tr>
<tr>
<td>3</td>
<td>\left( \frac{35}{31}, \frac{24}{31}, \frac{3}{31}, 0 \right)</td>
<td>-7.16</td>
<td></td>
<td></td>
<td></td>
<td>(0, 0, 0, 1)</td>
</tr>
</tbody>
</table>
10.7.2 Theorem Convex-simplex method of Zangwill.

Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) be continuously differentiable, and consider the problem to minimize \( f(x) \) subject to \( Ax = b, \ x \geq 0 \). Here \( A \) is an \( m \times n \) matrix and \( b \) is an \( m \)-vector such that all extreme points of the feasible region have \( m \) positive components and every choice of \( m \) columns of \( A \) is linearly independent. Suppose that the sequence \( \{x_k\} \) is generated by the convex-simplex method. Then any accumulation point is a KKT point.

**Proof**

Let \( \{x_k\}_H \) be a convergent subsequence with limit \( \hat{x} \). We need to show that \( \hat{x} \) is a KKT point. Suppose, by contradiction, that \( \hat{x} \) is not a KKT point. We shall exhibit a sequence \( \{(x_k, d_k)\}_H^* \) satisfying conditions 1 through 4 of Lemma 10.2.6, which is impossible.

Let \( \{d_k\}_H \) be the sequence of directions associated with \( \{x_k\}_H \). Note that \( d_k \) is defined by (10.46) through (10.55) at \( x_k \). Letting \( I_k \) be the set denoting the indices of the \( m \) largest components of \( x_k \) used to compute \( d_k \), there exists \( H' \subseteq H \) such that \( I_k = \hat{I} \) for each \( k \in H' \), where \( \hat{I} \) is the set denoting the indices of the \( m \) largest components of \( \hat{x} \). Furthermore, there exists \( H'' \subseteq H' \) such that \( d_k \) is given either by (10.53) and (10.55) for all \( k \in H'' \) or by (10.54) and (10.55) for all \( k \in H'' \). In the first case, let \( \hat{d} \) be obtained.
from (10.46), (10.47), (10.48), (10.49), (10.51), (10.53), and (10.55) at \( \hat{x} \), and in the latter case, let \( \hat{d} \) be obtained from (10.46), (10.47), (10.48), (10.50), (10.52), (10.54), and (10.55) at \( \hat{x} \). In either case, \( d_k = \hat{d} \) for \( k \in \mathcal{K}^* \). By the continuous differentiability of \( f \), note that \( \hat{d} \) would have been obtained by applying (10.46) through (10.55) at \( x_i \). By assumption, \( \hat{x} \) is not a KKT point, and hence \( \hat{d} \neq 0 \) and \( \nabla f(\hat{x})' \hat{d} < 0 \). To summarize, we have exhibited a sequence \( \{(x_k, d_k)\}_{x_i} \) satisfying Conditions 1, 2, and 4 of Lemma 10.2.6. To complete the proof we need to show that Part 3 also holds true.

Note that \( d_k = \hat{d} \) for \( k \in \mathcal{K}^* \). If \( \hat{d} \geq 0 \), then \( x_k + \lambda \hat{d} \geq 0 \) for all \( \lambda \in [0, \infty) \). If \( \hat{d} \not\geq 0 \), and since \( \hat{d} \) is a feasible direction at \( \hat{x} \), then \( \hat{x} + \lambda \hat{d} \geq 0 \) for all \( \lambda \in [0, 2\delta] \), where \( 2\delta = \min \{|\hat{x}_i|: \hat{d}_i < 0\} \). Since \( x_{ik} \rightarrow \hat{x}_i \) and \( d_k = \hat{d} \), then \( \delta_k = \min \{|-x_{ik}/d_{ik}: d_{ik} < 0|\} \geq \delta \) for all sufficiently large \( k \) in \( \mathcal{K}^* \). From (10.56) it then follows that \( x_k + \lambda d_k \) is feasible for all \( \lambda \in [0, \delta] \) for large \( k \) in \( \mathcal{K}^* \). Thus, Condition 3 of Lemma 10.2.6 holds true, and the proof is complete.

10.8 Effective First- and Second-Order Variants of the Reduced Gradient Method

In both the reduced gradient method of Wolfe and the convex-simplex method of Zangwill, we have seen how, given a feasible solution, we can partition the space into a set of basic variables \( x_B \) and a set of nonbasic variables \( x_N \) and then essentially project the problem onto the space of the nonbasic variables by substituting \( x_B = B^{-1}b - B^{-1}Nx_N \) (see Exercise 10.52). In this space the problem under consideration becomes the following, treating \( x_B \geq 0 \) as slack variables in the transformed constraints.

Minimize \( \{F(x_N) = f(B^{-1}b - B^{-1}Nx_N, x_N): B^{-1}Nx_N \leq B^{-1}b, x_N \geq 0\} \). (10.57)

Note that

\[
[\nabla F(x_N)]' = \left[ \frac{\partial f}{\partial x_N} + \frac{\partial f}{\partial x_B} \frac{\partial x_B}{\partial x_N} \right]' = \nabla_N f(x)' - \nabla_B f(x)' B^{-1}N = r_N',
\]

(10.58)

where \( r_N \) is the reduced gradient. Moreover, barring degeneracy, the only binding constraints are those nonnegativity constraints from the set \( x_N \geq 0 \) for which \( x_{ij} = 0 \) currently. Hence, the reduced gradient method constructs the direction-finding subproblem in the nonbasic variable space as follows, where \( J_N \) denotes the index set for the nonbasic variables:
Methods of Feasible Directions

Minimize \( \{ \nabla F(x_N) \} d_N = r_Nd_N = \sum_{j \in J_N} r_jd_j : -x_j|r_j| \leq d_j \leq |r_j| \) 
for all \( j \in J_N \). \hfill (10.59)

Observe that the trivial solution to (10.59) is to let \( d_j = |r_j| = -r_j \) if \( r_j \leq 0 \), and to let \( d_j = -x_j|r_j| = -x_jr_j \) if \( r_j > 0 \). This gives \( d_N \) as in (10.43); and then, using \( Ad = Bd_B + Nd_N = 0 \), we compute \( d_B = -B^{-1}Nd_N \) as in (10.44) to derive the direction of movement \( d' = (d_B', d_N') \).

The convex-simplex method examines the same direction-finding problem (10.59) but permits only one component of \( d_N \) to be nonzero, namely, the one that has the largest absolute value in the solution of (10.59). This gives the scaled unit vector \( d_N' \), and then \( d_B \) is calculated as before according to \( d_B = -B^{-1}Nd_N \) to produce \( d \). Hence, whereas the convex-simplex method changes only one component of \( d_N \), moving in a direction parallel to one of the polyhedron edges, or axes in the nonbasic variable space, the reduced gradient method permits all components to change as desired according to (10.59). It turns out that while the former strategy is unduly restrictive, the latter strategy also results in a slow progress, as blocking occurs upon taking short steps because of the many components that are changing simultaneously.

Computationally, a compromise between the two foregoing extremes has been found to be beneficial. Toward this end, suppose that we further partition the \( n - m \) variables \( x_N \) into \( (x_S, x_{N'}) \) and, accordingly, partition \( N \) into \([S, N']\).

The variables \( x_S \), indexed by \( J_S \), say, where \( 0 \leq |J_S| = s \leq n - m \), are called superbasic variables, and are usually chosen as a subset of the variables \( x_N \) that are positive (or strictly between lower and upper bounds when both types of bounds are specified in the problem). The remaining variables \( x_{N'} \) are still referred to as nonbasic variables. The idea is, therefore, to hold the variables \( x_{N'} \) fixed and to permit the variables \( x_S \) to be the “driving force” in guiding the iterates toward improving feasible points, with the basic variables \( x_B \) following suit as usual. Hence, writing \( d' = (d'_B, d'_S, d'_{N'}) \), we have \( d_{N'} = 0 \); and from \( Ad = 0 \), we get \( Bd_B + Sd_S = 0 \), or \( d_B = -B^{-1}Sd_S \). Accordingly, we get

\[
\begin{bmatrix}
  d_B \\
  d_S \\
  d_{N'}
\end{bmatrix} =
\begin{bmatrix}
  -B^{-1}S \\
  I \\
  0
\end{bmatrix} d_S = Zd_S,
\hfill (10.60)
\]

where \( Z \) is defined appropriately as an \( n \times s \) matrix. Problem (10.59) then reduces to the following direction-finding problem:
Minimize \( \{ \nabla f(x)^T d = \nabla f(x)^T Zd_S = [\nabla S f(x)^T - \nabla_B f(x)^T B^{-1} S]d_S \} \)

subject to \( -x_j |r_j| \leq d_j \leq r_j \) for all \( j \in J_S \).

(10.61)

Similar to Problem (10.59), the solution to (10.61) yields \( d_j = |r_j| = -r_j \) if \( r_j \leq 0 \), and \( d_j = -x_j |r_j| = -x_j r_j \) if \( r_j > 0 \), for all \( j \in J_S \). This gives \( d_S \), and then we obtain \( d \) from (10.60). Note that for the reduced gradient method, we had \( S = N \), that is, \( s = n - m \), whereas for the convex-simplex method, we had \( s = 1 \).

A recommended practical implementation using the foregoing concept proceeds as follows. (The commercial package MINOS adopts this strategy.) To initialize, based on the magnitude of the components \( d_j, j \in J_N \), in the solution to (10.59), some \( s \) components \( d_S \) of \( d \) are permitted to change independently. (MINOS simply uses the bounds \(-|r_j| \leq d_j \leq |r_j|, \forall j \in J_N\), in this problem.) This results in some set of \( s \) (positive) superbasic variables. The idea now is to execute the reduced gradient method in the space of the \((x_B, x_S)\) variables, holding \( x_N \) fixed, and using (10.61) as the direction-finding problem. Accordingly, this technique is sometimes referred to as a suboptimization strategy. However, during these iterations, if any component of \( x_B \) or \( x_S \) hits its bound of zero, it is transferred into the nonbasic variable set. Also, a pivot operation is performed only when a basic variable blocks at its bound of zero. (Hence, the method does not necessarily maintain the \( m \) most positive components as basic.) Upon pivoting, the basic variable is exchanged for a superbasic variable to give a revised basis, and the leaving basic variable is transferred into the nonbasic set. Noting that we always have \( x_S > 0 \), this process continues until either \( J_S = \emptyset \) (\( s = 0 \)), or \( \|r_S\| \leq \varepsilon \), where \( \varepsilon > 0 \) is some tolerance value. At this point, the procedure enters a pricing phase in which the entire vector \( r_N \) is computed. If the KKT conditions are satisfied within an acceptable tolerance, then the procedure stops. Otherwise, an additional variable, or a set of (significantly enterable) additional variables under the option of multiple pricing, are transferred from the nonbasic into the superbasic variable set, and the procedure continues. Because of the suboptimization feature, this strategy turns out to be computationally desirable, particularly for large-scale problems that contain several more variables than constraints.

**Second-Order Functional Approximations**

The direction-finding problem (10.61) adopts a linear approximation to the objective function \( f \). As we know, with steep ridged contours of \( f \), this can be prone to a slow zigzagging convergence behavior. Figure 10.21 illustrates this...
phenomenon in the context of the convex simplex method \((s = 1)\). The reduced gradient method \((s = 2)\) would zigzag in a similar fashion, behaving like an unconstrained steepest descent method until some constraint blocks this path (see Exercise 10.41). On the other hand, if we were to adopt a second-order approximation for \(f\) in the direction-finding problem, we can hope to accelerate the convergence behavior.

For example, if the function illustrated in Figure 10.21 were itself quadratic, then such a direction at the origin (using \(s = 2\)) would point toward its unconstrained minimum as shown by the dashed line in Figure 10.21. This would lead us directly to the point where this dashed line intersects the plane \(x_5 = 0\), whence, with \(s = 1\) now \((x_5\) being nonbasic), the next iteration would converge to the optimum (see Exercise 10.41).

The development of such a quadratic direction-finding problem is straightforward. At the current point \(x\), we minimize a second-order approximation to \(f(x + d)\) given by \(f(x) + \nabla f(x)'d + \frac{1}{2}d'\mathbf{H}(x)d\) over the linear manifold \(Ad = 0\), where \(d' = (d'_p, d'_S, d'_{N'})\), with \(d'_{N'} = 0\). This gives \(d = Zd_S\), as in (10.60), and therefore, the direction-finding problem is given as follows, where we have used (10.61) to write \(\nabla f(x)'d = \nabla f(x)'dZd_S = r'_Sd_S\).

![Figure 10.21 Zigzagging of the convex-simplex method.](image-url)
Minimize \( \left\{ r_S' d_S + \frac{1}{2} d_S'[Z' H(x) Z] d_S : d_S \in \mathbb{R}^s \right\} \). \( (10.62) \)

Note that \((10.62)\) represents an unconstrained minimization of the quadratic approximation to the objective function projected onto the space of the superbasic direction components. Accordingly, the \( s \times s \) matrix \( Z' H(x) Z \) is called the \textit{projected Hessian matrix} and can be dense even if \( H \) is sparse. However, hopefully, \( s \) is small (see Exercise 10.55). Setting the gradient of the objective function in \((10.62)\) equal to zero, we get

\[ [Z' H(x) Z] d_S = -r_S. \] \( (10.63) \)

Note that \( d_S = 0 \) solves \((10.63)\) if and only if \( r_S = 0 \). Otherwise, assuming that \( Z' H(x) Z \) is positive definite [which would be the case, for example, if \( H(x) \) is positive definite, since \( Z \) has full column rank], we have \( d_S = -[Z' H(x) Z]^{-1} r_S \neq 0 \); and moreover, from \((10.61)\) and \((10.63)\), \( \nabla f(x)' d = \nabla f(x)' Z d_S = r_S' d_S = -d_S'[Z' H(x) Z] d_S < 0 \), so that \( d = Z d_S \) is an improving feasible direction. Using this Newton-based direction, we can now perform a line search and proceed using the above suboptimization scheme, with \((10.62)\) replacing \((10.61)\) for the direction-finding step.

Practically, even if the Hessian \( H \) is available and is positive definite, we would most likely not be able to afford to use it exactly as described above. Typically, one maintains a positive definite approximation to the projected Hessian \( Z' H(x) Z \) that is updated from one iteration to the next using a quasi-Newton scheme. Note that \( H(x) \) or \( Z' H(x) Z \) are never actually computed, and only a Cholesky factorization \( LL' \) of the foregoing quasi-Newton approximation is maintained, while accounting for the variation in the dimension of the superbasic variables (see Murtagh and Saunders [1978]). Also, \( Z \) is never computed, but rather, an LU factorization of \( B \) is adopted. This factored form of \( B \) is used in the solution of the system \( \pi B = \nabla_B f(x)' \), from which \( r_S \) is computed via \( r_S' = \nabla_S f(x)' - \pi S \) as in \((10.61)\), as well as for the solution of \( d_B \) from the system \( Bd_B = -S d_S \) as in \((10.60)\), once \( d_S \) is determined.

For problems in which \( s \) can get fairly large (\( \geq 200 \), say) even a quasi-Newton approach becomes prohibitive. In such a case, a conjugate gradient approach becomes indispensable. Here, the conjugate gradient scheme is applied directly to the projected problem of minimizing \( F(d_S) = f(x + Zd_S) \). Note that in this projected space, \( \nabla F(d_S) = (\delta f/\delta x_B)(\partial x_B/\partial d_S) + (\delta f/\delta x_S) (\partial x_S/\partial d_S) + (\delta f/\delta x_N')(\partial x_N'/\partial d_S) = [-\nabla_B f(x)' B^{-1} S + \nabla_S f(x)' I + 0]' = r_S'. \) Hence, the direction \( d_S \) is taken as \(-r_S + \alpha d_S'\), where \( d_S' \) is the previous direction and \( \alpha \) is a multiplier determined by the particular conjugate gradient scheme. Under
appropriate conditions, either the quasi-Newton or the conjugate gradient approach leads to a superlinearly convergent process.

**Exercises**

[10.1] Solve the following problem by the Topkis-Veinott method starting from the point \((1, 3)\).

Minimize \(3(1-x_1)^2 - 10(x_2 - x_1^2)^2 + 2x_1^2 - 2x_1x_2 + e^{-2x_1 - x_2}\)
subject to \(2x_1^2 + x_2^2 \leq 16\)
\((x_2 - x_1)^2 + x_1 \leq 6\)
\(2x_1 + x_2 \geq 5.\)

[10.2] Consider the following problem:

Minimize \(2(x_1 - 3)^2 + (x_2 - 2)^2\)
subject to \(2x_1 - x_2 \leq 0\)
\(x_1 - 2x_2 + 3 = 0.\)

Starting from \(x = (1, 2)^T\), solve the problem by Zoutendijk’s procedure using the following two normalization methods:

a. \(|d_j| \leq 1\) for \(j = 1, 2\).

b. \(d'd \leq 1\).

[10.3] For each of the following cases, give a suitable characterization of the set of feasible directions at a point \(x \in S\):

\(S = \{x: Ax = b, x \geq 0\}\);

\(S = \{x: Ax \leq b, Qx = q, x \geq 0\}\);

\(S = \{x: Ax \geq b, x \geq 0\}\).

[10.4] Consider the following problem with lower and upper bounds on the variables:

Minimize \(f(x)\)
subject to \(a_j \leq x_j \leq b_j\) for \(j = 1, \ldots, n\).

Let \(x\) be a feasible point. Let \(\nabla_j = \frac{\partial f(x)}{\partial x_j}\), and consider Zoutendijk’s procedure for generating an improving feasible direction.

a. Show that an optimal solution to the direction-finding problem, using the normalization constraint \(|d_j| \leq 1\), is given by
\[
\begin{align*}
d_j = \begin{cases} 
-1 & \text{if } x_j > a_j \text{ and } \nabla_j \geq 0 \\
1 & \text{if } x_j < b_j \text{ and } \nabla_j < 0 \\
0 & \text{otherwise.}
\end{cases}
\end{align*}
\]

b. Show that an optimal solution to the direction-finding problem, using the normalization constraint \(d'd \leq 1\), is given by

\[
d_j = \begin{cases} 
-\frac{\nabla_j}{\left( \sum_{i \in I} \nabla_i^2 \right)^{1/2}} & \text{if } j \in I \\
0 & \text{if } j \notin I
\end{cases}
\]

where \(I = \{j : x_j > a_j \text{ and } \nabla_j \geq 0, \text{ or else } x_j < b_j \text{ and } \nabla_j < 0\}\).

c. Using the methods in Parts a and b, solve the following problem starting from the point \((-2, -3)\), and compare the trajectories obtained:

Minimize \(3x_1^2 - 2x_1x_2 + 4x_2^2 - 4x_1 - 3x_2\)

subject to \(-2 \leq x_1 \leq 0\)

\(-3 \leq x_2 \leq 1\).

d. Show that the direction-finding maps in both Parts a and b are not closed.

e. Prove convergence or give a counterexample showing that the feasible direction algorithms using the direction-finding procedures discussed in parts a and b do not converge to a KKT point.

[10.5] Solve the following problem by Zoutendijk’s method for linear constraints:

Minimize \(3x_1^2 + 2x_1x_2 + 2x_2^2 - 4x_1 - 3x_2 - 10x_3\)

subject to \(x_1 + 2x_2 + x_3 = 8\)

\(-2x_1 + x_2 \leq 1\)

\(x_1, x_2, x_3 \geq 0\).

[10.6] In Zoutendijk’s procedure, the following problem is solved to generate an improving feasible direction where \(I = \{i : g_i(x) = 0\}\):

Minimize \(z\)

subject to \(\nabla f(x)'d \leq z\)

\(\nabla g_i(x)'d \leq z\) \quad \text{for } i \in I\)

\(-1 \leq d_j \leq 1\) \quad \text{for } j = 1, \ldots, n\).
a. Show that the method cannot accommodate nonlinear equality constraints of the form \( h_i(x) = 0 \) by replacing each constraint by \( h_i(x) \leq 0 \) and \(-h_i(x) \leq 0\).

b. One way to handle a constraint of the form \( h_i(x) = 0 \) is to first replace it by the two constraints \( h_j(x) \leq \epsilon \) and \(-h_j(x) \leq \epsilon\), where \( \epsilon > 0 \) is a small scalar, and then to apply the above direction-finding process. Use this method to solve the following problem, starting from the point \((2, 1, 1)\).

\[
\begin{align*}
\text{Minimize} & \quad 3x_1^3 + 2x_2^2x_3 + 2x_3 \\
\text{subject to} & \quad x_1^2 + 2x_2 + x_3^2 = 7 \\
& \quad 2x_1^2 - 3x_2 + 2x_3 \leq 7.
\end{align*}
\]

[10.7] Consider the following problem, and let \( \hat{x} \) be a feasible point with \( g_i(\hat{x}) = 0 \) for \( i \in I:\)

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0 \quad \text{for} \quad i = 1, \ldots, m \\
& \quad h_i(x) = 0 \quad \text{for} \quad i = 1, \ldots, \ell.
\end{align*}
\]

a. Show that \( \hat{x} \) is a KKT point if and only if the optimal objective of the following problem is zero:

\[
\begin{align*}
\text{Minimize} & \quad \nabla f(\hat{x})'d \\
\text{subject to} & \quad \nabla g_i(\hat{x})'d \leq 0 \quad \text{for} \quad i \in I \\
& \quad \nabla h_i(\hat{x})'d = 0 \quad \text{for} \quad i = 1, \ldots, \ell \\
& \quad -1 \leq d_j \leq 1 \quad \text{for} \quad j = 1, \ldots, n.
\end{align*}
\]

b. Let \( \hat{d} \) be an optimal solution to the problem in Part a. If \( \nabla f(\hat{x})'\hat{d} < 0 \), then \( \hat{d} \) is an improving direction. Even though \( \hat{d} \) may not be a feasible direction, it is at least tangential to the feasible region at \( \hat{x} \).

The following procedure is proposed. Fix \( \delta > 0 \) and let \( \hat{\lambda} \) be an optimal solution to the problem to minimize \( f(\hat{x} + \lambda\hat{d}) \) subject to \( 0 \leq \lambda \leq \delta \). Let \( \overline{x} = \hat{x} + \hat{\lambda}\hat{d} \). Starting from \( \overline{x} \), a correction move is used to obtain a feasible point. This could be done in several ways.

1. Move along \( d = -(A'\lambda)^{-1}A'F(\overline{x}) \), where \( F \) is a vector function whose components are \( h_i(\overline{x}) \) for \( i = 1, \ldots, \ell \), and \( g_i(\overline{x}) \) for \( i \) such that \( g_i(\overline{x}) > 0 \), and \( A \) is the matrix whose rows are the
transposes of the gradients of the constraints in \( F \) (assumed to be linearly independent).

2. Use a penalty function scheme to minimize total infeasibility starting from \( \hat{x} \).

Using each of the above approaches, solve the problem given in Part b of Exercise 10.6.

[10.8] Solve the following problem using Zoutendijk's method for nonlinear constraints, starting from the point \((1, 3, 1)\).

Minimize \( 3x_1^2 + 2x_1x_2 + 2x_2^2 - 4x_1 - 3x_2 - 10x_3 \)
subject to \( x_1^2 + 2x_2^2 \leq 19 \)
\(-2x_1 + 2x_2 + x_3 \leq 5 \)
\( x_1, x_2, x_3 \geq 0. \)

[10.9] Consider the problem to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \). Suppose that \( x \) is a feasible point with \( g_i(x) = 0 \) for \( i \in I \). Furthermore, suppose that \( g_i \) is pseudoconcave at \( x \) for each \( i \in I \). Show that the following problem produces an improving feasible direction or concludes that \( x \) is a KKT point:

Minimize \( \nabla f(x)'d \)
subject to \( \nabla g_i(x)'d \leq 0 \) for \( i \in I \)
\(-1 \leq d_j \leq 1 \) for \( j = 1, \ldots, n \).

[10.10] In Section 10.1, in reference to Zoutendijk's method for linear constraints, we described several normalization constraints such as \( d'd \leq 1, -1 \leq d_j \leq 1 \) for \( j = 1, \ldots, n \), and \( \nabla f(x_k)'d \geq -1 \). Show that each of the following normalization constraints could be used instead:

a. \( \sum_{j=1}^{n} |d_j| \leq 1 \).

b. \( \max_{1 \leq j \leq n} |d_j| \leq 1 \).

c. \( A(x_k + d) \leq b \), provided that the set \( \{ x : Ax \leq b \} \) is bounded.

d. \( d_j \geq -1 \) if \( \frac{\partial f(x_k)}{\partial x_j} > 0 \) and \( d_j \leq 1 \) if \( \frac{\partial f(x_k)}{\partial x_j} < 0 \).

[10.11] Consider the following problem having linear constraints and nonlinear inequality constraints:
Minimize \( f(x) \)
subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \)
\( Ax \leq b \)
\( Qx = q. \)

Let \( x \) be a feasible point, and let \( I = \{i : g_i(x) = 0\} \). Furthermore, suppose that \( A_1x = b_1 \) and \( A_2x < b_2 \), where \( A' = [A'_1, A'_2] \) and \( b' = (b'_1, b'_2) \).

a. Show that the following linear program provides an improving feasible direction or concludes that \( x \) is a Fritz John point:

Minimize \( z \)
subject to \( \nabla f(x)'d - z \leq 0 \)
\( \nabla g_i(x)'d - z \leq 0 \) for \( i \in I \)
\( A_1d \leq 0 \)
\( Qd = 0. \)

b. Using this approach, solve the problem in Example 10.1.8, and compare the trajectories generated in both cases.

[10.12] Consider the project to minimize \( f(x) \) subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m \). Let \( \hat{x} \) be a feasible solution, and let \( I = \{i : g_i(\hat{x}) = 0\} \). Let \( (\hat{z}, \hat{d}) \) be an optimal solution to the following problem:

Minimize \( z \)
subject to \( \nabla f(\hat{x})'d \leq z \)
\( \nabla g_i(\hat{x})'d \leq z \) for \( i \in I \)
\( d_j \geq -1 \) if \( \frac{\partial f(\hat{x})}{\partial x_j} > 0 \)
\( d_j \leq 1 \) if \( \frac{\partial f(\hat{x})}{\partial x_j} < 0. \)

a. Show that \( \hat{z} = 0 \) if and only if \( \hat{x} \) is a Fritz John point.
b. Show that if \( \hat{z} < 0 \), then \( \hat{d} \) is an improving feasible direction.
c. Show how one of the binding constraints, instead of the objective function, could be used to bound the components of the vector \( d \).

[10.13] Consider the following problem:

Minimize \( f(x) \)
subject to \( g_i(x) \leq 0 \) for \( i = 1, \ldots, m. \)
The following is a modification of the direction-finding problem of Topkis and Veinott if \( g_i \) is pseudoconcave:

\[
\begin{align*}
\text{Minimize} & \quad \nabla f(x)^T d \\
\text{subject to} & \quad g_i(x) + \nabla g_i(x)^T d \leq 0 \quad \text{for } i = 1, \ldots, m \\
& \quad d^T d \leq 1.
\end{align*}
\]

a. Show that \( x \) is a KKT point if and only if the optimal objective value is equal to zero.

b. Let \( d \) be an optimal solution and suppose that \( \nabla f(x)^T \hat{d} < 0 \). Show that \( \hat{d} \) is an improving feasible direction.

c. Can you show convergence to a KKT point of the above modified Topkis and Veinott algorithm?

d. Repeat Parts a through c if the normalization constraint is replaced by \(-1 \leq d_j \leq 1 \) for \( j = 1, \ldots, n \).

e. Using the above approach, solve the problem in Example 10.1.5.

[10.14] Consider the following problem with lower and upper bounds on the variables:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad a_j \leq x_j \leq b_j \quad \text{for } j = 1, \ldots, n.
\end{align*}
\]

Let \( x \) be a feasible solution, let \( \nabla_j = \frac{\partial f(x)}{\partial x_j} \), and consider the modified Topkis-Veinott method for generating an improving feasible direction described in Exercise 10.13.

a. Show that an optimal solution to the direction-finding problem using the normalization constraint \( |d_j| \leq 1 \), is given by

\[
d_j = \begin{cases} 
\max\{a_j - x_j, -1\} & \text{if } \nabla_j \geq 0 \\
\min\{b_j - x_j, 1\} & \text{if } \nabla_j < 0.
\end{cases}
\]

b. Show that an optimal solution to the direction-finding problem, using the normalization constraint \( d^T d \leq \delta \) is given by

\[
d_j = \begin{cases} 
\max\{-\nabla_j/\|\nabla f(x)\|, a_j - x_j\} & \text{if } \nabla_j \geq 0 \\
\min\{-\nabla_j/\|\nabla f(x)\|, b_j - x_j\} & \text{if } \nabla_j < 0,
\end{cases}
\]

where
Methods of Feasible Directions

\[ \delta = \sum_{j: \nabla f_j \geq 0} \left[ \max\{-\nabla f_j / \|\nabla f(x)\|, a_j - x_j\} \right]^2 + \sum_{j: \nabla f_j < 0} \left[ \min\{-\nabla f_j / \|\nabla f(x)\|, b_j - x_j\} \right]^2. \]

c. Solve the problem in Part c of Exercise 10.4 by the methods in parts a and b above, and compare the trajectories obtained.

d. For the direction-finding maps in Parts a and b, show convergence of the method described above to a KKT point.

[10.15] Consider the problem to minimize \( f(x) \) subject to \( Ax \leq b \), where the region \( \{ x : Ax \leq b \} \) is bounded. Suppose that \( x_k \) is a feasible point, and let \( y_k \) solve the problem to minimize \( \nabla f(x_k)^t y \) subject to \( Ay \leq b \). Let \( \lambda_k \) be an optimal solution to the problem to minimize \( f[\lambda x_k + (1 - \lambda)y_k] \) subject to \( 0 \leq \lambda \leq 1 \), and let \( x_{k+1} = \lambda_k x_k + (1 - \lambda_k) y_k \).

a. Show that this procedure can be interpreted as a feasible direction method. Furthermore, show that, in general, the direction \( y_k - x_k \) cannot be obtained by Problems P1, P2, or P3 discussed in Section 10.1. Discuss any advantages or disadvantages of the above procedure.

b. Solve the problem given in Example 10.1.5 by the above method.

c. Describe the above procedure as the composition of a direction-finding map and a line search map. Using Theorem 7.3.2, show that the composite map is closed. Then, using Theorem 7.2.3, show that the algorithm converges to a KKT point.

d. Compare this method with the successive linear programming algorithm presented in Section 10.3.

The above procedure is credited to Frank and Wolfe [1956].)

[10.16] Consider the problem to minimize \( f(x) = e^t x + (1/2)x^tHx \) subject to \( Ax \leq b \). At a point \( x_k \) in the interior of the feasible region, Zoutendijk’s procedure of Section 10.1 generates a direction of movement by solving the problem to minimize \( \nabla f(x_k)^t d \) subject to \(-1 \leq d_j \leq 1 \) for \( j = 1, ..., n \). In Chapter 8 we learned that at interior points where we have an essentially unconstrained problem, conjugate direction methods are effective. The procedure discussed below combines a conjugate direction method with Zoutendijk’s method of feasible directions.

**Initialization Step**  Find an initial feasible solution \( x_1 \) with \( Ax_1 \leq b \). Let \( k = 1 \) and go to the Main Step.

**Main Step**
1. Starting from $x_k$, apply Zoutendijk's method, yielding $z$. If $Az < b$, let $y_1 = x_k$, $y_2 = z$, $\hat{d}_1 = y_2 - y_1$, $v = 2$, and go to Step 2. Otherwise, let $x_{k+1} = z$, replace $k$ by $k + 1$, and repeat Step 1.

2. Let $\hat{d}_v$ be an optimal solution to the following program:

$$\begin{align*}
\text{Minimize} & \quad \nabla f(y_v)'d \\
\text{subject to} & \quad \hat{d}_i' H d = 0 \quad \text{for } i = 1, \ldots, v-1 \\
& \quad -1 \leq d_j \leq 1 \quad \text{for } j = 1, \ldots, v.
\end{align*}$$

Let $\lambda_v$ be an optimal solution to the following line search problem:

$$\begin{align*}
\text{Minimize} & \quad f(y_v + \lambda \hat{d}_v) \\
\text{subject to} & \quad 0 \leq \lambda \leq \lambda_{\text{max}},
\end{align*}$$

where $\lambda_{\text{max}}$ is determined according to (10.1). Let $y_{v+1} = y_v + \lambda_v \hat{d}_v$.

If $Ay_{v+1} < b$ and $v \leq n - 1$, replace $v$ by $v + 1$, and repeat Step 2. Otherwise, replace $k$ by $k + 1$, let $x_k = y_{v+1}$, and go to Step 1.

a. Solve the problem in Exercise 10.14 by the procedure discussed above.

b. Using the above procedure, solve the following problem credited to Kunzi et al. [1966], starting from the point $(0, 0)$:

$$\begin{align*}
\text{Minimize} & \quad \frac{1}{2} x_1^2 + \frac{1}{2} x_2^2 - x_1 - 2x_2 \\
\text{subject to} & \quad 2x_1 + 3x_2 \leq 6 \\
& \quad x_1 + 4x_2 \leq 5 \\
& \quad x_1, x_2 \geq 0.
\end{align*}$$

c. Solve the problem in Parts a and b by replacing the Zoutendijk procedure at Step 1 of the above algorithm by the modified Topkis-Veinott algorithm discussed in Exercise 10.15.

d. Solve the problem in Parts a and b by replacing Zoutendijk's procedure at Step 1 of the above algorithm by the gradient projection method.

[10.17] Using the proof of Theorem 4.4.2 and Theorem 9.3.1 (see also Exercise 9.13), construct a detailed proof of Part a of Theorem 10.3.1.

[10.18] Derive analogues of the PSLP algorithm described in Section 10.3 by employing (a) a quadratic penalty function and (b) an augmented Lagrangian penalty function in lieu of the exact penalty function. Discuss the applicability, merits, and demerits of the procedures derived.
[10.19] Consider Problem P to minimize $f(x)$ subject to $Ax = b$ and $x \geq 0$. Suppose that $\bar{x}$ is feasible with $\bar{x}_j = 0$ for $j \in J_0$ and $\bar{x}_j > 0$ for $j \in J_+$. Also, assume that the Hessian $H(\bar{x})$ is positive definite.

a. Construct a problem for finding a direction $d$ that minimizes a second-order approximation of $f(\bar{x} + d)$ over the set of feasible directions at $\bar{x}$, with $\|d\|_\infty \leq 1$.

b. Suppose that $d = 0$ solves the problem of Part a. Show that $\bar{x}$ is then a KKT point for P.

c. Suppose that $d = 0$ is not an optimum for the problem of Part a. Show that the optimum to this problem then yields an improving feasible direction for P.

[10.20] Consider the problem of Example 10.3.2. Solve the associated KKT conditions to obtain the optimal Lagrange multipliers and, hence, prescribe a suitable value of $\mu$ to use in Algorithm PSLP. Also, find the eigenvalues and vectors of the Hessian of the objective function, along with the unconstrained minimum and, hence, sketch the contours of the objective function as in Figure 10.13a. State a suitable termination criterion for Algorithm PSLP and, using the starting iteration solution of Example 10.3.2, continue the algorithm until this termination criterion is satisfied.

[10.21] In view of Exercise 9.33, consider the linear programming Problem P to minimize $c'x$ subject to $Ax = 0$, $e'x = 1$, and $x \geq 0$, where $A$ is an $m \times n$ matrix of rank $m$ and $e$ is a vector of $n$ ones. Defining $Y = \text{diag}\{y_1, \ldots, y_n\}$, this problem can be written as:

Minimize $\{c'Y^2e : Ay^2e = 0$ and $e'Y^2e = 1\}$, where $x = Y^2e$.

Consider the following algorithm to solve this problem:

**Initialization** Select a feasible solution $x_0 > 0$, put $k = 0$, let $\delta \in (0, \sqrt{(n-1)/n})$, and go to the Main Step.

**Main Step** Let $y_k = \sqrt{x_k}$ for $j = 1, \ldots, n$ and define $Y_k = \text{diag}\{y_{k1}, \ldots, y_{kn}\}$. Solve the following subproblem (in $y$):

SP: Minimize $\{c'Y_ky : Ay_ky = 0, e'Y_ky = 1, \text{ and } \|Y_k^{-1}(y - y_k)\| \leq \delta\}$.

Let $y_{k+1}$ solve SP. Put $x_{k+1} = Y_k(2y_{k+1} - y_k)$, and let $x_{k+1} = x_{k+1}/e'x_{k+1}$. Increment $k$ by 1 and repeat the Main Step until a suitable convergence criterion holds true.

a. Interpret the derivation of SP in light of an SLP approach to solve Problem P. State a suitable termination criterion. Illustrate by
solving the problem to minimize \(-2x_1 + x_2 - x_3\) subject to \(2x_1 + 2x_2 - 3x_3 = 0, \quad x_1 + x_2 + x_3 = 1, \) and \(x \geq 0.\)

b. Note that in the linearization \(\alpha' (y - y_k) = 0\) of the constraint \(e'Y^2e = 1,\) we have used \(\alpha\) equal to the constraint gradient. Suppose, instead, that we use \(\alpha = Y_k^{-1}e.\) Solve the example of Part a for the resulting subproblem. (Morshed and Tapia [1987] show that the procedure of Part b is equivalent to Karmarkar's [1984] algorithm, and that the procedure of Part a is equivalent to the affine scaling variant of this algorithm.)

[10.22] Consider the system (10.21) and assume that \(\nabla^2 L(x_k)\) is positive definite and that the Jacobian \(\nabla h(x_k)\) has full row rank. Find an explicit closed-form solution \((d, v)\) to this system.

[10.23] Relate to an SQP approach the method of Roos and Vial [1988] as described in Exercise 9.34.

[10.24] Consider Problem P to minimize \(x_1 + x_2\) subject to \(x_1^2 + x_2^2 = 2.\) Find an optimal primal and dual solution to this problem. Now, consider the quadratic program \(QP(x_k, v_k)\) defined by (10.22) for any \((x_k, v_k),\) but with the Hessian \(\nabla^2 f(x_k)\) of the objective function incorrectly replacing the Hessian \(\nabla^2 L(x_k)\) of the Lagrangian. Comment on the outcome of doing this for the given Problem P. Now, starting at the point \(x = (1, 1)',\) apply the SQP approach to solve P using a suitable starting Lagrange multiplier \(v.\) What happens if \(v = 0\) is chosen as a starting value?

[10.25] Referring to Example 10.4.3, complete its solution using algorithm RSQP. Comment on its convergence behavior. Also, verify the optimality of the iterate \(x_2\) generated by algorithm MSQP using the corresponding quadratic programming subproblem.

[10.26] Let P: Minimize \(\{f(x): g_i(x) \leq 0, \quad i = 1,\ldots, m\}\) and consider the following quadratic programming direction-finding problem, where \(B_k\) is some positive definite approximation to the Hessian of the Lagrangian at \(x = x_k,\) and where \(\mu\) is large enough, as in Lemma 10.4.1.

\[
\text{QP: Minimize } \nabla f(x_k)'d + \frac{1}{2}d'B_kd + \mu \sum_{i=1}^{m} z_i \\
\text{subject to } z_i \geq g_i(x_k) + \nabla g_i(x_k)'d \quad \text{for } i = 1,\ldots, m \\
z_1,\ldots,z_m \geq 0.
\]

a. Discuss, in regard to feasibility, the advantage of Problem QP over (10.25). What is the accompanying disadvantage?
b. Let \( \mathbf{d}_k \) solve QP with optimum Lagrange multipliers \( \mathbf{u}_k \) associated with the first \( m \) constraints. If \( \mathbf{d}_k = \mathbf{0} \), then do we have a KKT solution for \( \mathbf{P} \)? Discuss.

c. Suppose that \( \mathbf{d}_k \neq \mathbf{0} \) solves QP. Show that \( \mathbf{d}_k \) is then a descent direction for the \( \ell_1 \) penalty function \( F_{\ell_1}(\mathbf{x}) = f(\mathbf{x}) + \mu \sum_{i=1}^{m} \max\{0, g_i(\mathbf{x})\} \) at \( \mathbf{x} = \mathbf{x}_k \).

d. Extend the above analysis to include equality constraints \( h_i(\mathbf{x}) = 0, i = 1, \ldots, \ell \), by considering the following subproblem QP:

\[
\begin{align*}
\text{Minimize} & \quad \nabla f(\mathbf{x}_k)^T \mathbf{d} + \frac{1}{2} \mathbf{d}^T \mathbf{B}_k \mathbf{d} + \mu \left[ \sum_{i=1}^{m} y_i + \sum_{i=1}^{\ell} (z_i^+ + z_i^-) \right] \\
\text{subject to} & \quad y_i \geq g_i(\mathbf{x}_k) + \nabla g_i(\mathbf{x}_k)^T \mathbf{d} \quad \text{for} \ i = 1, \ldots, m \\
& \quad z_i^+ - z_i^- \geq h_i(\mathbf{x}_k) + \nabla h_i(\mathbf{x}_k)^T \mathbf{d} \quad \text{for} \ i = 1, \ldots, \ell \\
& \quad \mathbf{y} \geq \mathbf{0}, z^+ \geq \mathbf{0}, z^- \geq \mathbf{0}.
\end{align*}
\]

[10.27] Provide a detailed proof for Theorem 10.4.2, defining precisely the input and output quantities associated with each algorithmic map, and supporting all the arguments in the proof sketched in Section 10.4.

[10.28] Consider the problem to minimize \( f(\mathbf{x}) \) subject to \( g_i(\mathbf{x}) \leq 0 \) for \( i = 1, \ldots, m \). The feasible direction methods discussed in this chapter start with a feasible point. This exercise describes a method for obtaining such a point if one is not immediately available. Select an arbitrary point \( \mathbf{x} \), and suppose that \( g_i(\mathbf{x}) \leq 0 \) for \( i \in I \) and \( g_i(\mathbf{x}) > 0 \) for \( i \notin I \). Now, consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad \sum_{i \notin I} y_i \\
\text{subject to} & \quad g_i(\mathbf{x}) \leq 0 \quad \text{for} \ i \in I \\
& \quad g_i(\mathbf{x}) - y_i \leq 0 \quad \text{for} \ i \notin I \\
& \quad y_i \geq 0 \quad \text{for} \ i \notin I.
\end{align*}
\]

a. Show that a feasible solution to the original problem exists if and only if the optimal objective value of the above problem is zero.

b. Let \( \mathbf{y} \) be a vector whose components are \( y_i \) for \( i \notin I \). The above problem could be solved by a feasible direction method starting from the point \( (\hat{\mathbf{x}}, \hat{\mathbf{y}}) \), where \( \hat{y}_i = g_i(\hat{\mathbf{x}}) \) for \( i \notin I \). At termination, a feasible solution to the original problem is obtained. Starting from this point, a feasible direction method could be used to solve the original problem. Illustrate this method by solving the following problem starting from the infeasible point \( (1, 2) \):
Minimize \( 2e^{-3x_1 - x_2} + x_1 x_2 + 2x_2^2 \)
subject to \( 3e^{-x_1} + x_2^2 \leq 4 \)
\( 2x_1 + 3x_2 \leq 6. \)

[10.29] Consider the following problem, where \( A \) is a \( v \times n \) matrix:

Minimize \( \nabla f(x)'d \)
subject to \( A_1d = 0 \)
\( d'd \leq 1. \)

The KKT conditions are both necessary and sufficient for optimality since a suitable constraint qualification holds true at each feasible solution (see Exercise 5.20). In particular, \( \bar{d} \) is an optimal solution if and only if there exist \( u \) and \( \mu \) such that

\[-\nabla f(x) = 2\mu \bar{d} + A_1'u \]
\[A_1\bar{d} = 0, \quad \bar{d}'\bar{d} \leq 1 \]
\[(\bar{d}'\bar{d} - 1)\mu = 0, \quad \mu \geq 0. \]

a. Show that \( \mu = 0 \) if and only if \( -\nabla f(x) \) is in the range-space of \( A_1' \) or, equivalently, if and only if the projection of \( -\nabla f(x) \) onto the nullspace of \( A_1 \) is the zero vector. In this case, \( \nabla f(x)'\bar{d} = 0. \)

b. Show that if \( \mu > 0 \), then an optimal solution \( \bar{d} \) to the above problem points in the direction of the projection of \( -\nabla f(x) \) onto the nullspace of \( A_1 \).

c. Show that a solution to the KKT system stated above could be immediately obtained as follows. Let \( u = -(A_1A_1')^{-1}A_1\nabla f(x), \) and let \( d = -(I - A_1'(A_1A_1')^{-1}A_1)\nabla f(x) \). If \( d = 0 \), let \( \mu = 0 \) and \( \bar{d} = 0. \) If \( d \neq 0 \), let \( \mu = \|d\|/2 \) and \( \bar{d} = d/\|d\|. \)

d. Now, consider the problem to minimize \( f(x) \) subject to \( Ax \leq b. \) Let \( x \) be a feasible solution such that \( A_1x = b_1 \) and \( A_2x < b_2, \) where \( A' = (A_1', A_2') \) and \( b' = (b_1', b_2'). \) Show that \( x \) is a KKT point of the problem to minimize \( f(x) \) subject to \( Ax \leq b \) if \( \mu = 0 \) and \( u \geq 0. \)

e. Show that if \( \mu = 0 \) and \( u \geq 0, \) then the gradient projection method discussed in Section 10.5 continues by choosing a negative component \( u_j \) of \( u, \) deleting the associated row from \( A_1, \) producing \( A_1' \) and resolving the direction-finding problem to minimize
Methods of Feasible Directions

\[ \nabla f(x)^t d \text{ subject to } A_1^t d = 0, \ d^t d \leq 1. \] Show that the optimal solution to this problem is not equal to zero.

f. Solve the problem of Example 10.5.5 by the gradient projection method, where the projected gradient is found by minimizing \[ \nabla f(x)^t d \text{ subject to } A_1 d = 0, \ d^t d \leq 1. \]

[10.30] Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad A x \leq b.
\end{align*}
\]

Let x be a feasible solution and let \( A_1 x = b_1, \) \( A_2 x \leq b_2, \) where \( A' = (A_1', A_2') \) and \( b' = (b_1', b_2'). \) Zoutendijk's method of feasible directions finds a direction by solving the problem to minimize \( \nabla f(x)^t d \) subject to \( A_1 d \leq 0 \) and \( d^t d \leq 1. \) In view of Exercise 10.29, the gradient projection method finds a direction by solving the problem to minimize \( \nabla f(x)^t d \) subject to \( A_1 d = 0 \) and \( d^t d \leq 1. \)

a. Compare the methods, pointing out their advantages and disadvantages. Also, compare these methods with the successive linear programming algorithm described in Section 10.3.

b. Solve the following problem starting from \( (0,0) \) by the method of Zoutendijk, the gradient projection method, and Algorithm PSLP of Section 10.3, and compare their trajectories:

\[
\begin{align*}
\text{Minimize} & \quad 3x_1^2 + 2x_1x_2 + 2x_2^2 - 6x_1 - 9x_2 \\
\text{subject to} & \quad -3x_1 + 6x_2 \leq 9 \\
& \quad -2x_1 + x_2 \leq 1 \\
& \quad x_1, x_2 \geq 0.
\end{align*}
\]

[10.31] Solve the following problem by the gradient projection method:

\[
\begin{align*}
\text{Minimize} & \quad (2 - x_1)^2 - 8(x_2 - x_1)^2 + 2x_1^2 - 2x_1x_2 + e^{-2x_1} - x_2 \\
\text{subject to} & \quad 5x_1 + 6x_2 \leq 30 \\
& \quad -4x_1 + 3x_2 \leq 12 \\
& \quad x_1, x_2 \geq 0.
\end{align*}
\]

[10.32] Consider the constraints \( A x \leq b \) and let \( P = I - A_1^t (A_1 A_1^t)^{-1} A_1, \) where \( A_1 \) represents the gradients of the binding constraints at a given feasible point \( \hat{x}. \) What are the implications and geometric interpretation of the following three statements?

a. \( P \nabla f(\hat{x}) = 0. \)

b. \( P \nabla f(\hat{x}) = \nabla f(\hat{x}). \)
Chapter 10

10.33 Consider the following problem, where $A_1$ is a $v \times n$ matrix:

Minimize $\| -\nabla f(x) - d \|^2$

subject to $A_1d = 0$.

c. $P\nabla f(\hat{x}) \neq 0$.

[10.34] Consider the problem to minimize $f(x)$ subject to $Ax \leq b$. The following modification to Zoutendijk’s method and to the gradient projection method of Rosen is proposed. Given a feasible point $x$, if $-\nabla f(x)$ is feasible, then the direction of movement $d$ is taken as $-\nabla f(x)$; otherwise, the direction $d$ is computed according to the respective algorithms.

a. Using the above modification, solve the problem in Example 10.5.5 starting from $x_1 = (0.0, 0.0)'$ by Rosen’s gradient projection method. Compare the trajectory with that obtained in Example 10.5.5.

b. Using the above modification, solve the problem in Example 10.1.5 starting from $x_1 = (0, 0.75)'$ by Zoutendijk’s method. Compare the trajectory with that obtained in Example 10.1.5.

[10.35] Consider the following problem:

Minimize $2x_1^2 + 3x_2^2 + 3x_3^2 + 2x_1x_2 - 2x_1x_3 + x_2x_3 - 5x_1 - 3x_2$

subject to $3x_1 + 2x_2 + x_3 \leq 6$

$a. Solve the problem by Zoutendijk’s method of feasible directions, starting from the origin.

b. Solve the problem by the gradient projection method, starting from the origin.
[10.36] Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad c^t x \\
\text{subject to} & \quad Ax = b \\
& \quad x \geq 0,
\end{align*}
\]

where \( A \) is an \( m \times n \) matrix of rank \( m \). Consider solving the problem by the gradient projection method.

a. Let \( x \) be a basic feasible solution and let \( d = -Pc \), where \( P \) projects any vector onto the nullspace of the gradients of the binding constraints. Show that \( d = 0 \).

b. Let \( u = -(MM^t)^{-1}Mc \), where the rows of \( M \) are the transposes of the gradients of the binding constraints. Show that deleting the row corresponding to the most negative \( u_j \) associated with the constraint \( x_j \geq 0 \), forming a new projection matrix \( P' \), and moving along the direction \(-P'e\) is equivalent to entering the variable \( x_j \) in the basis in the simplex method.

c. Using the results of parts a and b, show that the gradient projection method reduces to the simplex method if the objective function is linear.

[10.37] Consider the following problem, where \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) is differentiable:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad -x \leq 0.
\end{align*}
\]

a. Suppose that \( x \) is a feasible solution, and suppose that \( x' = (x_1', x_2') \), where \( x_1 = 0 \) and \( x_2 > 0 \). Denote \( \nabla f(x)' \) by \( (\nabla_1', \nabla_2') \). Show that the direction \( d' \) generated by the gradient projection method is given by \((0, -\nabla_2')\).

b. If \( \nabla_2 = 0 \), show that the gradient projection method simplifies as follows. If \( \nabla_1 \geq 0 \), stop; \( x \) is a KKT point. Otherwise, let \( j \) be any index such that \( x_j = 0 \) and \( \partial f(x)/\partial x_j < 0 \). Then the new direction of movement \( d = (0, \ldots, 0, -\partial f(x)/\partial x_j, 0, \ldots, 0)' \), where \( \partial f(x)/\partial x_j \) appears at position \( j \).

c. Illustrate the method by solving the following problem:

\[
\begin{align*}
\text{Minimize} & \quad 3x_1^2 + 2x_1x_2 + 4x_2^2 + 5x_1 + 3x_2 \\
\text{subject to} & \quad x_1, x_2 \geq 0.
\end{align*}
\]
d. Solve the problem in Example 10.2.3 starting from the point \((0, 0.1, 0)\) using the above procedure.

[10.38] In the gradient projection method, if \(\nabla f(x) = 0\), we throw away from the matrix \(A_1\) a row corresponding to a negative component of the vector \(u\). Suppose that, instead, we throw away all rows corresponding to negative components of the vector \(u\). Show by a numerical example that the resulting projection matrix does not necessarily lead to an improving feasible direction.

[10.39] In the gradient projection method, we often calculate \((A_1A_1')^{-1}\) in order to compute the projection matrix. Usually, \(A_1\) is updated by deleting or adding a row to \(A_1\). Rather than computing \((A_1A_1')^{-1}\) from scratch, the old \((A_1A_1')^{-1}\) could be used to compute the new \((A_1A_1')^{-1}\).

a. Suppose that \(C = \begin{pmatrix} C_1 & C_2 \\ C_3 & C_4 \end{pmatrix} \) and \(C^{-1} = \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix} \). Show that \(C_1^{-1} = B_1 - B_2B_4^{-1}B_3\). Furthermore, suppose that \(C_1^{-1}\) is known, and show that \(C^{-1}\) could be computed by letting

\[
\begin{align*}
B_1 &= C_1^{-1} + C_1^{-1}C_2C_0^{-1}C_3C_1^{-1} \\
B_2 &= -C_1^{-1}C_2C_0^{-1} \\
B_3 &= -C_0^{-1}C_3C_1^{-1} \\
B_4 &= C_0^{-1}
\end{align*}
\]

where \(C_0 = C_4 - C_2C_1^{-1}C_2\).

b. Simplify the above formulas for the gradient projection method, both when adding and deleting a row. (In the gradient projection method, \(C_1 = A_1A_1'\), \(C_2 = A_1a\), \(C_3 = a'A_1\), and \(C_4 = a'a\), where \(a'\) is the row added when \(C_1^{-1}\) is known, or deleted when \(C^{-1}\) is known.)

c. Use the gradient projection method, with the scheme described in this exercise for updating \((A_1A_1')^{-1}\), to solve the following problem starting from the solution \((2, 1, 3)\).

Minimize \(3x_1^2 + 2x_1x_2 + 2x_2^2 + 2x_1^2 + 2x_2x_3 + 3x_1 + 5x_2 + 8x_3\)

subject to \(x_1 + x_2 + x_3 \leq 6\)

\[-x_1 - x_2 + 2x_3 \geq 3\]

\(x_1, x_2, x_3 \geq 0.\)
[10.40] In the reduced gradient method, suppose that the set $I_k$ defined in (10.40) consists of indices of any $m$ positive variables. Investigate if the direction-finding map is closed.

[10.41] a. Consider the problem illustrated in Figure 10.21, and assume that the convex-simplex method of Section 10.7 is used to solve this problem. Use the graph to illustrate a plausible path followed by the algorithm, specifying the set of basic and nonbasic variables, the signs on the components of the reduced gradient $r_N$, and the result of the line search at each iteration.

b. Repeat to illustrate a conceivable trajectory for the reduced gradient algorithm of Section 10.6.

c. Repeat to illustrate the effect of using the quadratic programming subproblem (10.62), starting at the origin with $x_1$ and $x_2$ as super-basic variables, and assuming that the objective function is itself quadratic.

[10.42] Modify the rules of the convex-simplex method such that it handles directly the problem to minimize $f(x)$ subject to $Ax = c$, $a \leq x \leq b$. Use the method to solve the following problem:

Minimize $2e^{-x_1} + 3x_1^2 - x_1x_2 + 2x_2^2 + 6x_1 - 5x_2$

subject to $\begin{align*}
3x_1 + 2x_2 &\leq 12 \\
-2x_1 + 3x_2 &\leq 6 \\
1 &\leq x_1, x_2 \leq 3.
\end{align*}$

[10.43] Show that the direction-finding map of the convex-simplex method defined by (10.46) through (10.55) is closed.

[10.44] Consider the following problem:

Minimize $x_1^3 + 2x_2^3 + 3x_1 - 4x_2 - x_1^2$

subject to $\begin{align*}
3x_1 + 2x_2 &\leq 6 \\
-x_1 + 2x_2 &\leq 4 \\
x_1, x_2 &\geq 0.
\end{align*}$

Solve the problem by the convex-simplex method. Is the solution obtained a global optimum, a local optimum, or neither?

[10.45] Consider the following problem:

Minimize $x_1^2 + x_1x_2 + 2x_2^2 - 6x_1 - 14x_2$

subject to $\begin{align*}
x_1 + x_2 + x_3 &= 4 \\
-x_1 + 3x_2 &\leq 1 \\
x_1, x_2, x_3 &\geq 0.
\end{align*}$
Using the starting solution (2, 1, 1):

a. Solve the problem by the gradient projection method.
b. Solve the problem by the reduced gradient method.
c. Solve the problem by the convex-simplex method.
d. Solve the problem by the PSLP algorithm.
e. Solve the problem by the MSQP algorithm.

[10.46] Consider the following problem:

\[
\text{Minimize } f(x) \\
\text{subject to } \begin{align*}
Ax &= b \\
x &\geq 0.
\end{align*}
\]

Assume that \( f \) is a concave function and that the feasible region is compact, so that, by Theorem 3.4.7, an optimal extreme point exists.

a. Show how the convex-simplex method could be modified so that it only searches among extreme points of the feasible region.
b. At termination, a KKT point is at hand. Is this point necessarily an optimal solution, a local optimal solution, or neither? If the point is not optimal, can you develop a cutting plane method that excludes the current point but not the optimal solution?
c. Illustrate the procedures in Parts a and b by solving the following problem, starting from the origin:

\[
\text{Minimize } -(x_1 - 2)^2 - (x_2 - 1)^2 \\
\text{subject to } \begin{align*}
-2x_1 + x_2 &\leq 4 \\
3x_1 + 2x_2 &\leq 12 \\
3x_1 - 2x_2 &\leq 6 \\
x_1, x_2 &\geq 0.
\end{align*}
\]

[10.47] As originally proposed, given a point \( x \), the reduced gradient method moves along the direction \( d \), where (10.43) is modified as follows:

\[
d_j = \begin{cases} 
-r_j & \text{if } x_j > 0 \text{ or } r_j \leq 0 \\
0 & \text{otherwise}
\end{cases}
\]

a. Prove that \( d = 0 \) if and only if \( x \) is a KKT point.
b. Show that if \( d \neq 0 \), then \( d \) is an improving feasible direction.
c. Using the above direction-finding map, solve the following problem by the reduced gradient method:

\[
\text{Minimize } 3e^{-2x_1+x_2} + 2x_1^2 + 2x_1x_2 + 3x_2^2 + x_1 + 3x_2 \\
\text{subject to } \begin{align*}
2x_1 + x_2 &\leq 4 \\
-x_1 + x_2 &\leq 3 \\
x_1, x_2 &\geq 0.
\end{align*}
\]
d. Show that the direction-finding map given above is not closed.

[10.48] Show that the convex-simplex method reduces to the simplex method if the objective function is linear.

[10.49] For both the reduced gradient method and the convex-simplex method it was assumed that each feasible solution has at least $m$ positive components. This exercise gives a necessary and sufficient condition for this to hold true. Consider the set $S = \{x : Ax = b, x \geq 0\}$, where $A$ is an $m \times n$ matrix of rank $m$. Show that each $x \in S$ has at least $m$ positive components if and only if every extreme point of $S$ has exactly $m$ positive components.

[10.50] Suppose that in the convex-simplex method, the direction-finding process is modified as follows. The scalar $\beta$ in (10.50) is computed as

$$
\beta = \begin{cases} 
\max \{r_j : x_j > 0 \text{ and } r_j \geq 0\} & \text{if } x_i > 0 \text{ and } r_i \geq 0 \text{ for some } i \\
0 & \text{otherwise.}
\end{cases}
$$

Furthermore, the index $v$ is computed as

$$
v = \begin{cases} 
\nu & \text{is an index such that } \alpha = -r_v \text{ if } \alpha \geq \beta \\
\nu & \text{is an index such that } \beta = r_v \text{ if } \alpha < \beta.
\end{cases}
$$

Show that with this modification, the direction-finding map is not necessarily closed.

[10.51] Consider the following problem:

Minimize $c^T x + \frac{1}{2} x^T H x$

subject to $Ax = b$

$x \geq 0$.

Suppose that the constraint $h(x) = Ax - b = 0$ is handled by a penalty function of the form $\mu h(x)^T h(x)$, giving the following problem:

Minimize $c^T x + \frac{1}{2} x^T H x + \mu (Ax - b)^T (Ax - b)$

subject to $x \geq 0$.

Give the detailed steps of a feasible direction method for solving the above problem. Illustrate the method for the following data:

$$
H = \begin{bmatrix}
2 & -2 & 0 \\
-2 & 3 & 0 \\
0 & 0 & 0
\end{bmatrix}, \quad A = \begin{bmatrix}
1 & 2 & 0 \\
2 & 1 & 2
\end{bmatrix}, \quad c = \begin{bmatrix}
-3 \\
-2 \\
0
\end{bmatrix}, \quad b = \begin{bmatrix}
4 \\
7
\end{bmatrix}.
$$
[10.52] Consider Problem P to minimize \( f(x) \) subject to \( Ax = b \) and \( x \geq 0 \), where \( A \) is an \( m \times n \) matrix of rank \( m \). Given a solution \( \bar{x} = (\bar{x}_B, \bar{x}_N) \), where \( \bar{x}_B > 0 \) and where the corresponding partition \((B, N)\) of \( A \) has \( B \) nonsingular, write the basic variables \( x_B \) in terms of the nonbasic variables \( x_N \) to derive the following representation of \( P \) in the nonbasic variable space:

\[
P(x_N) : \text{Minimize } \{ F(x_N) = f(B^{-1}b - B^{-1}Nx_N, x_N) : B^{-1}Nx_N \leq B^{-1}b, x_N \geq 0 \}.
\]

a. Identify the set of binding constraints for \( P(x_N) \) at the current solution \( \bar{x}_N \). Relate \( \nabla F(x_N) \) to the reduced gradient for Problem \( P \).

b. Write a set of necessary and sufficient conditions for \( \bar{x}_N \) to be a KKT point for \( P(x_N) \). Compare this with the result of Theorem 10.6.1.

[10.53] This exercise describes the method credited to Griffith and Stewart [1961] for solving a nonlinear programming problem by successively approximating it by a sequence of linear programs. Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad g(x) \leq 0 \\
& \quad h(x) = 0 \\
& \quad a \leq x \leq b,
\end{align*}
\]

where \( f: \mathbb{R}^n \to \mathbb{R} \), \( g: \mathbb{R}^n \to \mathbb{R}^m \), and \( h: \mathbb{R}^n \to \mathbb{R} \). Given the point \( x_k \) at iteration \( k \), the algorithm replaces \( f, g, \) and \( h \) by their linear approximations at \( x_k \), yielding the following linear programming problem, where \( \nabla g(x_k) \) is an \( m \times n \) matrix denoting the Jacobian of the vector function \( g \), and \( \nabla h(x_k) \) is an \( \ell \times n \) matrix denoting the Jacobian of the vector function \( h \):

\[
\begin{align*}
\text{Minimize} & \quad f(x_k) + \nabla f(x_k)'(x - x_k) \\
\text{subject to} & \quad g(x_k) + \nabla g(x_k)(x - x_k) \leq 0 \\
& \quad h(x_k) + \nabla h(x_k)(x - x_k) = 0 \\
& \quad a \leq x \leq b.
\end{align*}
\]

**Initialization Step** Choose a feasible point \( x_1 \), choose a parameter \( \delta > 0 \) that limits the movement at each iteration, and choose a termination scalar \( \varepsilon > 0 \). Let \( k = 1 \), and go to the Main Step.

**Main Step**

1. Solve the following linear programming problem:
Minimize \( \nabla f(x_k)'(x - x_k) \)
subject to \( \nabla g(x_k)(x - x_k) \leq -g(x_k) \)
\( \nabla h(x_k)(x - x_k) = -h(x_k) \)
\( a \leq x \leq b \)
\(-\delta \leq x_i - x_{ik} \leq \delta \)
for \( i = 1, \ldots, n \),

where \( x_{ik} \) is the \( i \)th component of \( x_k \). Let \( x_{k+1} \) be an optimal solution, and go to Step 2.

2. If \( \|x_{k+1} - x_k\| \leq \varepsilon \) and if \( x_{k+1} \) is near-feasible, stop with \( x_{k+1} \).
   Otherwise, replace \( k \) by \( k + 1 \) and go to Step 1.

Even though convergence of the above method is not generally guaranteed, the method is reported to be effective for solving many practical problems.

a. Construct an example that shows that if \( x_k \) is feasible to the original problem, then \( x_{k+1} \) is not necessarily feasible.

b. Now, suppose that \( h \) is linear. Show that if \( g \) is concave, the feasible region to the linear program is contained in the feasible region of the original problem. Furthermore, show that if \( g \) is convex, then the feasible region to the original problem is contained in the feasible region to the linear program.

c. Solve the following problem in Walsh [1975, p. 67] both by the method described in this exercise and by Kelley's cutting plane algorithm presented in Exercise 7.23, and compare their trajectories:

Minimize \(-2x_1^2 + x_1x_2 - 3x_2^2\)
subject to \(3x_1 + 4x_2 \leq 12\)
\(x_1^2 - x_2^2 \geq 1\)
\(0 \leq x_1 \leq 4\)
\(0 \leq x_2 \leq 3\).

d. Re-solve the example of Part c using the PSLP algorithm of Section 10.3 and compare the trajectories obtained.

[10.54] Consider the bilinear program to minimize \( \phi(x, y) = c'x + d'y + x'Hy \)
subject to \( x \in X \) and \( y \in Y \), where \( X \) and \( Y \) are bounded polyhedral sets in \( R^n \) and \( R^m \), respectively. Consider the following algorithm.

**Initialization Step** Select an \( x_1 \in R^n \) and \( y_1 \in R^m \). Let \( k = 1 \) and go to the Main Step.
Main Step

1. Solve the linear program to minimize $d^T y + x_k^T H y$ subject to $y \in Y$. Let $\hat{y}$ be an optimal solution. Let $y_{k+1}$ be as specified below and go to Step 2.

$$y_{k+1} = \begin{cases} y_k & \text{if } \phi(x_k, \hat{y}) = \phi(x_k, y_k) \\ \hat{y} & \text{if } \phi(x_k, \hat{y}) < \phi(x_k, y_k). \end{cases}$$

2. Solve the linear program to minimize $c^T x + x_k^T H y_{k+1}$ subject to $x \in X$. Let $\hat{x}$ be an optimal solution. Let $x_{k+1}$ be as specified below and go to Step 3.

$$x_{k+1} = \begin{cases} x_k & \text{if } \phi(\hat{x}, y_{k+1}) = \phi(x_k, y_{k+1}) \\ \hat{x} & \text{if } \phi(\hat{x}, y_{k+1}) < \phi(x_k, y_{k+1}). \end{cases}$$

3. If $x_{k+1} = x_k$ and $y_{k+1} = y_k$, stop with $(x_k, y_k)$ as a KKT point. Otherwise, replace $k$ by $k + 1$ and go to Step 1.

a. Using the above algorithm, find a KKT point to the bilinear program to minimize $2x_1y_1 + 3x_2y_2$, subject to $x \in X$ and $y \in Y$, where

$$X = \{x : x_1 + 3x_2 \geq 30, 2x_1 + x_2 \geq 20, 0 \leq x_1 \leq 27, 0 \leq x_2 \leq 16\}$$

$$Y = \left\{y : \frac{5}{3}y_1 + y_2 \geq 10, y_1 + y_2 \leq 15, 0 \leq y_1 \leq 10, 0 \leq y_2 \leq 10\right\}.$$

b. Prove that the algorithm converges to a KKT point.

c. Show that if $(x_k, y_k)$ is such that $\phi(x_k, y_k) \leq \min\{\phi(x_k', y) : y \in Y\}$ for all extreme points $x_k'$ of $X$ that are adjacent to $x_k$ (including $x_k$ itself), and if $\phi(x_k, y_k) \leq \min\{\phi(x, y_k') : x \in X\}$ for all extreme points $y_k'$ of $Y$ that are adjacent to $y_k$ (including $y_k$ itself), then $(x_k, y_k)$ is a local minimum for the bilinear program. (This result is discussed in Vaish [1974].)

[10.55] Consider the nonlinear programming Problem P to minimize $f(x)$ subject to $Ax = b$ and $x \geq 0$, where $A$ is an $m \times n$ matrix of rank $m$, and suppose that some $q \leq n$ variables appear nonlinearly in the problem. Assuming that P has an optimum, show that there exists an optimum in which the number of superbasic variables $s$ in a reduced gradient approach satisfies $s \leq q$. Extend this result to include nonlinear constraints.

[10.56] This exercise gives a generalization of Wolfe's reduced gradient method for handling nonlinear equality constraints. This generalized procedure was
developed by Abadie and Carpentier [1969], and a modified version is given below for brevity. Consider the following problem:

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad h_i(x) = 0 \quad \text{for } i = 1, \ldots, \ell \\
& \quad a_j \leq x_j \leq u_j \quad \text{for } j = 1, \ldots, n.
\end{align*}
\]

Here we assume that \(f\) and \(h_i\) for each \(i\) are differentiable. Let \(h\) be the vector function whose components are \(h_i\) for \(i = 1, \ldots, \ell\), and, furthermore, let \(a\) and \(u\) be the vectors whose components are \(a_j\) and \(u_j\) for \(j = 1, \ldots, n\). We make the following nondegeneracy assumption. Given any feasible solution \(x'\), it can be decomposed into \((x_B, x_N)\) with \(x_B \in \mathbb{R}^\ell\) and \(x_N \in \mathbb{R}^{n-\ell}\), where \(a_B < x_B < u_B\). Furthermore, the \(\ell \times n\) Jacobian matrix \(\nabla h(x)\) is decomposed accordingly into the \(\ell \times \ell\) matrix \(\nabla_B h(x)\) and the \(\ell \times (n-\ell)\) matrix \(\nabla_N h(x)\), such that \(\nabla_B h(x)\) is invertible. The following is an outline of the procedure.

**Initialization Step** Choose a feasible solution \(x'\) and decompose it into \((x'_B, x'_N)\). Go to the Main Step.

**Main Step**

1. Let \(r' = \nabla_N f(x') - \nabla_B f(x')^T \nabla_B h(x)^{-1} \nabla_N h(x)\). Compute the \((n - \ell)\) vector \(d_N\) whose \(j\)th component \(d_j\) is given by

\[
d_j = \begin{cases} 
0 & \text{if } x_j = a_j \text{ and } r_j > 0, \text{ or } x_j = u_j \text{ and } r_j < 0 \\
-r_j & \text{otherwise.}
\end{cases}
\]

If \(d_N = 0\), stop; \(x\) is a KKT point. Otherwise, go to Step 2.

2. Find a solution to the nonlinear system \(h(y, x_N) = 0\) by Newton's method as follows, where \(x_N\) is specified below.

**Initialization** Choose \(\varepsilon > 0\) and a positive integer \(K\). Let \(\theta > 0\) be such that \(a_N \leq x_N \leq u_N\), where \(x_N = x_N + \theta d_N\). Let \(y_1 = x_B\), let \(k = 1\), and go to Iteration \(k\) below.

**Iteration \(k\)**

(i) Let \(y_{k+1} = y_k - \nabla_B h(y_k, \bar{x}_N)^{-1} h(y_k, \bar{x}_N)\). If \(a_B \leq y_{k+1} \leq u_B\), \(f(y_{k+1}, \bar{x}_N) < f(x_B, x_N)\), and \(\|h(y_{k+1}, \bar{x}_N)\| < \varepsilon\), go to Step (iii); otherwise, go to Step (ii).
(ii) If \( k = K \), replace \( \theta \) by \((1/2)\theta\), let \( \bar{x}_N = x_N + \theta d_N \), let \( y_1 = x_B \), replace \( k \) by 1, and go to Step (i). Otherwise, replace \( k \) by \( k + 1 \) and go to Step (i).

(iii) Let \( x' = (y_{k+1}', \bar{x}_N') \), choose a new basis \( B \), and go to Step 1 of the main algorithm.

a. Using the above algorithm, solve the following problem:

Minimize \( 2x_1^2 + 2x_1x_2 + 3x_2^2 + 10x_1 - 2x_2 \)
subject to \( 2x_1^2 - x_2 = 0 \)
\( 1 \leq x_1, x_2 \leq 2. \)

b. Show how the procedure could be modified to handle inequality constraints as well. Illustrate by solving the following problem:

Minimize \( 2x_1^2 + 2x_1x_2 + 3x_2^2 + 10x_1 - 2x_2 \)
subject to \( x_1^2 + x_2^2 \leq 9 \)
\( 1 \leq x_1, x_2 \leq 2. \)

[10.57] In this exercise we describe a method credited to Davidon [1959] and developed later by Goldfarb [1969] for minimizing a quadratic function in the presence of linear constraints. The method extends the Davidon-Fletcher-Powell method and retains conjugacy of the search directions in the presence of constraints. Part e of the exercise also suggests an alternative approach. Consider the following problem:

Minimize \( c'x + \frac{1}{2} x'Hx \)
subject to \( Ax = b, \)

where \( H \) is an \( n \times n \) symmetric positive definite matrix and \( A \) is an \( m \times n \) matrix of rank \( m \). The following is a summary of the algorithm.

**Initialization Step** Let \( \varepsilon > 0 \) be a selected termination tolerance. Choose a feasible point \( x_1 \) and an initial symmetric positive definite matrix \( D_1 \). Let \( k = j = 1 \), let \( y_1 = x_1 \), and go to the Main Step.

**Main Step**

1. If \( \|\nabla f(y_j)\| < \varepsilon \), stop; otherwise, let \( d_j = -\hat{D}_j \nabla f(y_j) \), where

\[
\hat{D}_j = D_j - D_j A'(AD_j A')^{-1} AD_j.
\]

Let \( \lambda_j \) be an optimal solution to the problem to minimize \( f(y_j + \lambda d_j) \) subject to \( \lambda \geq 0 \), and let \( y_{j+1} = y_j + \lambda_j d_j \). If \( j < n \), go to Step
2. If $j = n$, let $y_1 = x_{k+1} = y_{n+1}$, replace $k$ by $k + 1$, let $j = 1$, and repeat Step 1.

2. Construct $D_{j+1}$ as follows:

$$D_{j+1} = D_j + \frac{p_j^T D_j q_j - D_j q_j}{p_j^T q_j} D_j q_j,$$

where $p_j = \lambda_j d_j$, and $q_j = \nabla f(y_{j+1}) - \nabla f(y_j)$. Replace $j$ by $j + 1$, and go to Step 1.

a. Show that the points generated by the algorithm are feasible.
b. Show that the search directions are $H$-conjugate.
c. Show that the algorithm stops in, at most, $n - m$ steps with an optimal solution.
d. Solve the following problem by the method described in this exercise:

Minimize $2x_1^2 + 3x_2^2 + 2x_3^2 + 3x_4^2 + 2x_1 x_2 - x_2 x_3 + 2x_3 x_4 - 3x_1 - 2x_2 + 5x_3$

subject to $3x_1 + x_2 + 2x_3 = 8$

$-2x_1 + x_2 + 3x_3 + x_4 = 6$.

e. Consider the following alternative approach. Decompose $x^T$ and $A$ into $(x_B^T, x_N^T)$ and $[B, N]$, respectively, where $B$ is an invertible $m \times m$ matrix. The system $Ax = b$ is equivalent to $x_B = B^{-1}b - B^{-1}Nx_N$. Substituting for $x_B$ in the objective function, a quadratic form involving the $(n - m)$-vector $x_N$ is obtained. The resulting function is then minimized using a suitable conjugate direction method such as the Davidon-Fletcher-Powell method. Use this approach to solve the problem given in Part d, and compare the two solution procedures.

f. Modify the above scheme, using the BFGS quasi-Newton update.

g. Extend both the methods discussed in this exercise for handling a general nonlinear objective function.

Notes and References

The method of feasible directions is a general concept that is exploited by primal algorithms that proceed from one feasible solution to another. In Section 10.1 we present the methods of Zoutendijk for generating improving feasible directions. It is well known that the algorithmic map used in Zoutendijk's method is not closed, and this is shown in Section 10.2. Furthermore, an example credited to Wolfe [1972] was presented, which shows that the
procedure does not generally converge to a KKT point. To overcome this difficulty, based on the work of Zoutendijk [1960], Zangwill [1969] presented a convergent algorithm using the concept of near-binding constraints. In Section 10.2 we describe another approach, credited to Topkis and Veinott [1967]. This method uses all the constraints, both active and inactive, and thereby avoids a sudden change in direction as a new constraint becomes active.

Note that the methods of unconstrained optimization discussed in Chapter 8 could be combined effectively with the method of feasible directions. In this case, an unconstrained optimization method is used at interior points, whereas feasible directions are generated at boundary points by one of the methods discussed in this chapter. An alternative approach is to place additional conditions at interior points, which guarantee that the direction generated is conjugate to some of the directions generated previously. This is illustrated in Exercise 10.16. Also refer to Kunzi et al. [1966], Zangwill [1967b], and Zoutendijk [1960]. Zangwill [1967a] developed a procedure for solving quadratic programming problems in a finite number of steps using the convex-simplex method in conjunction with conjugate directions.

In Sections 10.3 and 10.4 we present a very popular and effective class of successive linear programming (SLP) and successive quadratic programming (SQP) feasible direction approaches. Griffith and Stewart [1961] introduced the SLP approach at Shell as a method of approximation programming (see Exercise 10.53). Other similar approaches were developed by Buzby [1974] for a chemical process model at Union Carbide, and by Boddington and Randall [1979] for a blending and refinery problem at the Chevron Oil Company (see also Baker and Ventker [1980]). Beale [1978] describes a combination of SLP ideas with the reduced gradient method, and Palacios-Gomez et al. [1982] present an SLP approach using the $\ell_1$ penalty function as a merit function. Although intuitively appealing and popular because of the availability of efficient linear programming solvers, the foregoing methods are not guaranteed to converge. A first convergent form, which is described in Section 10.3 as the penalty SLP (PSLP) approach, was presented by Zhang et al. [1985] and uses the $\ell_1$ penalty function directly in the linear programming subproblem along with trust region ideas, following Fletcher [1981b]. Baker and Lasdon [1985] describe a simplified version of this algorithm that has been used to solve nonlinear refinery models having up to 1000 rows.

The SQP method concept (or the projected Lagrangian or the Lagrange-Newton method) was first used by Wilson [1963] in his SOLVER procedure as described by Beale [1967]. Han [1976] and Powell [1978b] suggest quasi-Newton approximations to the Hessian of the Lagrangian, and Powell [1978c] provides related superlinear convergence arguments. Han [1975b] and Powell [1978b] show how the $\ell_1$ penalty function can be used as a merit function to derive globally convergent variants of SQP. This is described as the MSQP method herein. See Crane et al. [1980] for a software description. For the use of other smooth exact penalty (ALAG) functions in this context and related discussion, see Fletcher [1987], Gill et al. [1981], and Schittkowski [1983]. The Maratos effect (Maratos [1978]) is described nicely in Powell [1986], and ways
to avoid it have been proposed by permitting increases in both the objective function and constraint violations (Chamberlain et al. [1982], Powell and Yuan [1986], and Schittkowski [1981]) or by altering search directions (Coleman and Conn [1982a,b] and Mayne and Polak [1982]), or through second-order corrections (Fletcher [1982b]) (see also Fukushima [1986]). For ways of handling infeasibility or unbounded quadratic subproblems, see Fletcher [1987] and Burke and Han [1989]. Fletcher [1981, 1987] describes the $L_1$SQP algorithm mentioned in Section 10.4 that combines the $\ell_1$ penalty function with trust region methods to yield a robust and very effective procedure. Tamura and Kobayashi [1991] describe experiences with an actual application of SQP, and Eldersveld [1991] discusses techniques for solving large-scale problems using SQP, along with a comprehensive discussion and computational results. For extensions of SQP methods to nonsmooth problems, see Pshenichnyi [1978] and Fletcher [1987].

In 1960, Rosen developed the gradient projection method for linear constraints and later, in 1961, generalized it for nonlinear constraints. Du and Zhang [1989] provide a comprehensive convergence analysis for a mild modification of this method as stated in Section 10.5. (An earlier analysis appears in Du and Zhang [1986].) In Exercises 10.29, 10.30, and 10.33, different methods for yielding the projected gradient are presented, and the relationship between Rosen's method and that of Zoutendijk is explored. In 1969, Goldfarb extended the Davidon-Fletcher-Powell method to handle problems having linear constraints utilizing the concept of gradient projection. In Exercise 10.57 we describe how equality constraints could be handled. For the inequality constrained problem, Goldfarb develops an active set approach that identifies a set of constraints that could be regarded as binding and applies the equality-constrained approach. The method was generalized by Davies [1970] to handle nonlinear constraints. Also refer to the work of Sargent and Murtagh [1973] on their variable metric projection method.

The method of reduced gradient was developed by Wolfe [1963] with the direction-finding map discussed in Exercise 10.47. In 1966, Wolfe provided an example to show that the method does not converge to a KKT point. The modified version described in Section 10.6 is credited to McCormick [1970a]. The generalized reduced gradient (GRG) method was later presented by Abadie and Carpentier [1969], who gave several approaches to handle nonlinear constraints. One such approach is discussed in Section 10.6, and another in Exercise 10.56. Computational experience with the reduced gradient method and its generalization is reported in Abadie and Carpentier [1967], Abadie and Guigon [1970], and Faure and Huard [1965]. Convergence proofs for the GRG method are presented under very restrictive and difficult to verify conditions. (See Smeers [1974, 1977], and Mokhtar-Kharroubi, [1980].) Improved versions of this method are presented in Abadie [1978a], Lasdon et al. [1978], and Lasdon and Waren [1978, 1982]. In Section 10.7 we discuss the convex-simplex method of Zangwill for solving a nonlinear programming problem having linear constraints. This method can be viewed as a reduced gradient method where only one nonbasic variable is changed at a time. A comparison of the convex-
simplex method with the reduced gradient method is given by Hans and Zangwill [1972].

In Section 10.8 we present the concept of superbasic variables, which unifies and extends the reduced gradient and convex-simplex methods, and we discuss the use of second-order functional approximations in the superbasic variable space to accelerate the algorithmic convergence. Murtagh and Saunders [1978] present a detailed analysis and algorithmic implementation techniques, along with appropriate factorization methods (see also Gill et al. [1981]). A description of the code MINOS is also presented here, as well as in Murtagh and Saunders [1982, 1983]. Shanno and Marsten [1982] show how the conjugate gradient method can be used to enhance the algorithm and present related restart schemes to maintain second-order information and obtain descent directions.

In this chapter we discuss search methods for solving a constrained nonlinear programming problem by generating feasible improving directions. Several authors have extended some of the unconstrained optimization techniques to handle simple constraints, such as linear constraints and lower and upper bounds on the variables. One way of handling constraints is to modify an unconstrained optimization technique simply by rejecting infeasible points during the search procedure. However, this approach is not effective, since it leads to premature termination at nonoptimal points. This was demonstrated by the results quoted by Friedman and Pinder [1972].

As we discussed earlier in these notes, Goldfarb [1969a] and Davies [1970] have extended the Davidon-Fletcher-Powell method to handle linear and nonlinear constraints, respectively. Several gradient-free search methods have also been extended to handle constraints. Glass and Cooper [1965] extended the method of Hooke and Jeeves to deal with constraints. Another attempt to modify the method of Hooke and Jeeves to accommodate constraints is credited to Klingman and Himmelblau [1964]. By projecting the search direction into the intersection of the binding constraints, Davies and Swann [1969] were able to incorporate linear constraints in the method of Rosenbrock with line searches. In Exercise 8.51 we described a variation of the simplex method of Spendley et al. [1962]. Box [1965] developed a constrained version of the simplex method. Several other alternative versions of this method were developed by Friedman and Pinder [1972], Ghani [1972], Guin [1968], Mitchell and Kaplan [1968], and Umida and Ichikawa [1971]. Another method that uses the simplex technique in constrained optimization was proposed by Dixon [1973a]. At interior points, the simplex method is used, with occasional quadratic approximations to the function. Whenever constraints are encountered, an attempt is made to move along the boundary. In 1973, Keefer proposed a method in which the basic search uses the Nelder and Mead simplex technique. The lower and upper bound constraints on the variables are dealt with explicitly, while other constraints are handled by a penalty function scheme.

Paviani and Himmelblau [1969] also use the simplex method in conjunction with a penalty function to handle constrained problems. The basic approach is to define a tolerance criterion \( \phi_k \) at iteration \( k \) and a penalty function \( P(x) \), as discussed in Chapter 9, so that the constraints can be replaced
by \( P(x) \leq \phi_k \). In the implementation of the Nelder and Mead method, a point is accepted only if it satisfies this criterion, and \( \phi_k \) is decreased at each iteration. Computational results using this approach are given by Himmelblau [1972b].

Several studies for evaluating and testing nonlinear programming algorithms have been made. Stocker [1969] compared five methods for solving 15 constrained and unconstrained optimization test problems of varying degrees of difficulty. In 1970, Colville made a comparative study of many nonlinear programming algorithms. In the study, 34 codes were tested by many participants, who attempted to solve eight test problems using their own methods and codes. A summary of the results of this study are reported in Colville [1970]. Computational results are also reported in Himmelblau [1972b]. Both studies use a range of nonlinear programming problems having varying degrees of difficulty, including highly nonlinear constraints and objective functions, linear constraints, and simple bounds on the variables. Discussions on the comparative performance and evaluation of various algorithms are given in Colville [1970] and Himmelblau [1972b]. For further software description and computational comparisons of reduced gradient-type methods, see Bard [1998], Gomes and Martinez [1991], Lasdon [1985], Waren et al. [1987], and Wasil et al. [1989].
Appendix A
Mathematical Review

In this appendix we review notation, basic definitions, and results related to vectors, matrices, and real analysis that are used throughout the text. For more details, see Bartle [1976], Berge [1963], Berge and Ghoulia-Houri [1965], Buck [1965], Cullen [1972], Flet [1966], and Rudin [1964].

A.1 Vectors and Matrices

Vectors

An n-vector \( \mathbf{x} \) is an array of \( n \) scalars \( x_1, x_2, \ldots, x_n \). Here \( x_j \) is called the \( j \)th component, or element, of the vector \( \mathbf{x} \). The notation \( \mathbf{x} \) represents a column vector, whereas the notation \( \mathbf{x}^t \) represents the transposed row vector. Vectors are denoted by lowercase boldface letters, such as \( \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{x}, \) and \( \mathbf{y} \). The collection of all \( n \)-vectors forms the \( n \)-dimensional Euclidean space, which is denoted by \( \mathbb{R}^n \).

Special Vectors

The zero vector, denoted by \( \mathbf{0} \), is a vector consisting entirely of zeros. The sum vector is denoted by \( \mathbf{1} \) or \( \mathbf{e} \) and has each component equal to 1. The \( i \)th coordinate vector, also referred to as the \( i \)th unit vector, is denoted by \( \mathbf{e}_i \) and consists of zeros except for a 1 at the \( i \)th position.

Vector Addition and Multiplication by a Scalar

Let \( \mathbf{x} \) and \( \mathbf{y} \) be two \( n \)-vectors. The sum of \( \mathbf{x} \) and \( \mathbf{y} \) is written as the vector \( \mathbf{x} + \mathbf{y} \). The \( j \)th component of the vector \( \mathbf{x} + \mathbf{y} \) is \( x_j + y_j \). The product of a vector \( \mathbf{x} \) and a scalar \( \alpha \) is denoted by \( \alpha \mathbf{x} \) and is obtained by multiplying each element of \( \mathbf{x} \) by \( \alpha \).

Linear and Affine Independence

A collection of vectors \( \mathbf{x}_1, \ldots, \mathbf{x}_k \) in \( \mathbb{R}^n \) is considered linearly independent if \( \sum_{j=1}^{k} \lambda_j \mathbf{x}_j = 0 \) implies that \( \lambda_j = 0 \) for all \( j = 1, \ldots, k \). A collection of vectors \( \mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k \) in \( \mathbb{R}^n \) is considered to be affinely independent if \( (\mathbf{x}_1 - \mathbf{x}_0), \ldots, (\mathbf{x}_k - \mathbf{x}_0) \) are linearly independent.
Linear, Affine, and Convex Combinations and Hulls

A vector $y$ in $\mathbb{R}^n$ is said to be a linear combination of the vectors $x_1, \ldots, x_k$ in $\mathbb{R}^n$ if $y$ can be written as $y = \sum_{j=1}^{k} \lambda_j x_j$ for some scalars $\lambda_1, \ldots, \lambda_k$. If, in addition, $\lambda_1, \ldots, \lambda_k$ are restricted to satisfy $\sum_{j=1}^{k} \lambda_j = 1$, then $y$ is said to be an affine combination of $x_1, \ldots, x_k$. Furthermore, if we also restrict $\lambda_1, \ldots, \lambda_k$ to be nonnegative, then this is known as a convex combination of $x_1, \ldots, x_k$. The linear, affine, or convex hull of a set $S \subseteq \mathbb{R}^n$ is, respectively, the set of all linear, affine, or convex combinations of points within $S$.

Spanning Vectors

A collection of vectors $x_1, \ldots, x_k$ in $\mathbb{R}^n$, where $k \geq n$, is said to span $\mathbb{R}^n$ if any vector in $\mathbb{R}^n$ can be represented as a linear combination of $x_1, \ldots, x_k$. The cone spanned by a collection of vectors $x_1, \ldots, x_k$, for any $k \geq 1$, is the set of nonnegative linear combinations of these vectors.

Basis

A collection of vectors $x_1, \ldots, x_k$ in $\mathbb{R}^n$ is called a basis of $\mathbb{R}^n$ if it spans $\mathbb{R}^n$ and if the deletion of any of the vectors prevents the remaining vectors from spanning $\mathbb{R}^n$. It can be shown that $x_1, \ldots, x_k$ form a basis of $\mathbb{R}^n$ if and only if $x_1, \ldots, x_k$ are linearly independent and if, in addition, $k = n$.

Inner Product

The inner product of two vectors $x$ and $y$ in $\mathbb{R}^n$ is defined by $x^t y = \sum_{j=1}^{n} x_j y_j$. If the inner product of two vectors is equal to zero, then the two vectors are said to be orthogonal.

Norm of a Vector

The norm of a vector $x$ in $\mathbb{R}^n$ is denoted by $\|x\|$ and defined by $\|x\| = (x^t x)^{1/2} = (\sum_{j=1}^{n} x_j^2)^{1/2}$. This is also referred to as the $\ell_2$ norm, or Euclidean norm.

Schwartz Inequality

Let $x$ and $y$ be two vectors in $\mathbb{R}^n$, and let $|x^t y|$ denote the absolute value of $x^t y$. Then the following inequality, referred to as the Schwartz inequality, holds true:

$$|x^t y| \leq \|x\| \|y\|.$$
Matrices

A matrix is a rectangular array of numbers. If the matrix has $m$ rows and $n$ columns, it is called an $m \times n$ matrix. Matrices are denoted by boldface capital letters, such as $A$, $B$, and $C$. The entry in row $i$ and column $j$ of a matrix $A$ is denoted by $a_{ij}$, its $i$th row is denoted by $A_i$, and its $j$th column is denoted by $a_j$.

Special Matrices

An $m \times n$ matrix whose elements are all equal to zero is called a zero matrix and is denoted by $0$. A square $n \times n$ matrix is called the identity matrix if $a_{ii} = 0$ for $i \neq j$ and $a_{ii} = 1$ for $i = 1, \ldots, n$. The $n \times n$ identity matrix is denoted by $I$ and sometimes by $I_n$ to highlight its dimension. An $n \times n$ permutation matrix $P$ is one that has the same rows of $I$, but which are permuted in some order. An orthogonal matrix $Q$ having dimension $m \times n$ is one that satisfies $Q'Q = I_n$ or $QQ' = I_m$. In particular, if $Q$ is square, $Q^{-1} = Q'$. Note that a permutation matrix $P$ is an orthogonal square matrix.

Addition of Matrices and Scalar Multiplication of a Matrix

Let $A$ and $B$ be two $m \times n$ matrices. The sum of $A$ and $B$, denoted by $A + B$, is the matrix whose $(i, j)$th entry is $a_{ij} + b_{ij}$. The product of a matrix $A$ by a scalar $\alpha$ is the matrix whose $(i, j)$th entry is $\alpha a_{ij}$.

Matrix Multiplication

Let $A$ be an $m \times n$ matrix and $B$ be an $n \times p$ matrix. Then the product $AB$ is defined to be the $m \times p$ matrix $C$ whose $(i, j)$th entry $c_{ij}$ is given by

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} \quad \text{for } i = 1, \ldots, m, \text{ and } j = 1, \ldots, p.$$ 

Transposition

Let $A$ be an $m \times n$ matrix. The transpose of $A$, denoted by $A^t$, is the $n \times m$ matrix whose $(i, j)$th entry is equal to $a_{ji}$. A square matrix $A$ is said to be symmetric if $A = A^t$. It is said to be skew symmetric if $A^t = -A$.

Partitioned Matrices

A matrix can be partitioned into submatrices. For example, the $m \times n$ matrix $A$ could be partitioned as follows:
Determinant of a Matrix

Let $A$ be an $n \times n$ matrix. The determinant of $A$, denoted by $\det[A]$, is defined iteratively as follows:

$$
\det[A] = \sum_{i=1}^{n} a_{i1} \det[A_{i1}].
$$

Here $A_{i1}$ is the cofactor of $a_{i1}$, defined as $(-1)^{i+1}$ times the submatrix of $A$ formed by deleting the $i$th row and the first column, and the determinant of any scalar is the scalar itself. Similar to the use of the first column above, the determinant can be expressed in terms of any row or column.

Inverse of a Matrix

A square matrix $A$ is said to be nonsingular if there is a matrix $A^{-1}$, called the inverse matrix, such that $AA^{-1} = A^{-1}A = I$. The inverse of a square matrix, if it exists, is unique. Furthermore, a square matrix has an inverse if and only if its determinant is not equal to zero.

Rank of a Matrix

Let $A$ be an $m \times n$ matrix. The rank of $A$ is the maximum number of linearly independent rows or, equivalently, the maximum number of linearly independent columns of the matrix $A$. If the rank of $A$ is equal to $\min\{m, n\}$, $A$ is said to have full rank.

Norm of a Matrix

Let $A$ be an $n \times n$ matrix. Most commonly, the norm of $A$, denoted by $\|A\|$, is defined by

$$
\|A\| = \max_{\|x\|=1} \|Ax\|
$$

where $\|Ax\|$ and $\|x\|$ are the usual Euclidean ($\ell_2$) norms of the corresponding vectors. Hence, for any vector $z$, $\|Az\| \leq \|A\| \|z\|$. A similar use of an $\ell_p$ norm $\|\cdot\|_p$ induces a corresponding matrix norm $\|A\|_p$. In particular, the above matrix norm, sometimes denoted $\|A\|_2$, is equal to the [maximum eigenvalue of $A^tA$]$^{1/2}$. Also, the Frobenius norm of $A$ is given by
\[ \|A\|_F = \left( \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2} \]

and is simply the \(\ell_2\) norm of the vector whose elements are all the elements of \(A\).

**Eigenvalues and Eigenvectors**

Let \(A\) be an \(n \times n\) matrix. A scalar \(\lambda\) and a nonzero vector \(x\) satisfying the equation \(Ax = \lambda x\) are called, respectively, an *eigenvalue* and an *eigenvector* of \(A\). To compute the eigenvalues of \(A\), we solve the equation \(\det(A - \lambda I) = 0\). This yields a polynomial equation in \(\lambda\) that can be solved for the eigenvalues of \(A\). If \(A\) is symmetric, then it has \(n\) (possibly nondistinct) eigenvalues. The eigenvectors associated with distinct eigenvalues are necessarily orthogonal, and for any collection of some \(p\) coincident eigenvalues, there exists a collection of \(p\) orthogonal eigenvectors. Hence, given a symmetric matrix \(A\), we can construct an orthogonal basis \(B\) for \(\mathbb{R}^n\), that is, a basis having orthogonal column vectors, each representing an eigenvector of \(A\). Furthermore, let us assume that each column of \(B\) has been normalized to have a unit norm. Hence, \(B'B = I\), so that \(B^{-1} = B'\). Such a matrix is said to be an *orthogonal matrix* or an *orthonormal matrix*.

Now, consider the (pure) quadratic form \(x'^t Ax\), where \(A\) is an \(n \times n\) symmetric matrix. Let \(\lambda_1, \ldots, \lambda_n\) be the eigenvalues of \(A\), let \(\Lambda = \text{diag} \{\lambda_1, \ldots, \lambda_n\}\) be a *diagonal matrix* comprised of diagonal elements \(\lambda_1, \ldots, \lambda_n\) and zeros elsewhere, and let \(B\) be the orthogonal eigenvector matrix comprised of the orthogonal, normalized eigenvectors \(b_1, \ldots, b_n\) as its columns. Define the linear transformation \(x = By\) that writes any vector \(x\) in terms of the eigenvectors of \(A\). Under this transformation, the given quadratic form becomes

\[ x'^t Ax = y'^t B'^t ABy = y'^t \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2. \]

This is called a *diagonalization process*.

Observe also that we have \(AB = BA\), so that because \(B\) is orthogonal, we get \(A = B \Lambda B' = \sum_{i=1}^{n} \lambda_i b_i b_i'^t\). This representation is called the *spectral decomposition* of \(A\). For an \(m \times n\) matrix \(A\), a related factorization \(A = U \Sigma V'^t\), where \(U\) is an \(m \times m\) orthogonal matrix, \(V\) is an \(n \times n\) orthogonal matrix, and \(\Sigma\) is an \(m \times n\) matrix having elements \(\Sigma_{ij} = 0\) for \(i \neq j\), and \(\Sigma_{ij} \geq 0\) for \(i = j\), is known as a *singular-value decomposition* (SVD) of \(A\). Here, the columns of \(U\) and \(V\) are normalized eigenvectors of \(AA'^t\) and \(A'^t A\), respectively. The \(\Sigma_{ij}\) values are the (absolute) square roots of the eigenvalues of \(AA'^t\) if \(m \leq n\) or of \(A'^t A\) if \(m \geq n\). The number of nonzero \(\Sigma_{ij}\) values equals the rank of \(A\).
Definite and Semidefinite Matrices

Let $A$ be an $n \times n$ symmetric matrix. Here $A$ is said to be positive definite if $x'Ax > 0$ for all nonzero $x$ in $\mathbb{R}^n$ and is said to be positive semidefinite if $x'Ax \geq 0$ for all $x$ in $\mathbb{R}^n$. Similarly, if $x'Ax < 0$ for all nonzero $x$ in $\mathbb{R}^n$, then $A$ is called negative definite; and if $x'Ax \leq 0$ for all $x$ in $\mathbb{R}^n$, then $A$ is called negative semidefinite. A matrix that is neither positive semidefinite nor negative semidefinite is called indefinite. By the foregoing diagonalization process, the matrix $A$ is positive definite, positive semidefinite, negative definite, and negative semidefinite if and only if its eigenvalues are positive, nonnegative, negative, and nonpositive, respectively. (Note that the superdiagonalization algorithm discussed in Chapter 3 is a more efficient method for ascertaining definiteness properties.) Also, by the definition of $A$ and $B$ above, if $A$ is positive definite, then its square root $A^{1/2}$ is the matrix satisfying $A^{1/2}A^{1/2} = A$ and is given by $A^{1/2} = BA^{1/2}B'$.

A.2 Matrix Factorizations

Let $B$ be a nonsingular $n \times n$ matrix, and consider the system of equations $Bx = b$. The solution given by $x = B^{-1}b$ is seldom computed by finding the inverse $B^{-1}$ directly. Instead, a factorization or decomposition of $B$ into multiplicative components is usually employed whereby $Bx = b$ is solved in a numerically stable fashion, often through the solution of triangular systems via back-substitution. This becomes particularly relevant in ill-conditioned situations when $B$ is nearly singular or when we wish to verify positive definiteness of $B$ as in quasi-Newton or Levenberg-Marquardt methods. Several useful factorizations are discussed below. For more details, including schemes for updating such factors in the context of iterative methods, we refer the reader to Bartels et al. [1970], Bazaraa et al. [2005], Dennis and Schnabel [1983], Dongarra et al. [1979], Gill et al. [1974, 1976], Golub and Van Loan [1983/1989], Murty [1983], and Stewart [1973], along with the many accompanying references cited therein. Standard software such as LINPACK, MATLAB, and the Harwell Library routines are also available to perform these factorizations efficiently.

LU and PLU Factorization for a Basis $B$

In the LU factorization, we reduce $B$ to an upper triangular form $U$ through a series of permutations and Gaussian pivot operations. At the $i$th stage of this process, having reduced $B$ to $B^{(i-1)}$, say, which is upper triangular in columns $1, \ldots, i - 1$ (where $B^0 = B$), we first premultiply $B^{(i-1)}$ by a permutation matrix $P_i$ to exchange row $i$ with that row in $\{i, i + 1, \ldots, n\}$ of $B^{(i-1)}$ that has the largest absolute-valued element in column $i$. This is done to ensure that the $(i, i)$th element of $P_iB^{(i-1)}$ is significantly nonzero. Using this as a pivot element, we
perform row operations to zero out the elements in rows \( i + 1, \ldots, n \) of column \( i \). This *triangularization* can be represented as a premultiplication with a suitable *Gaussian pivot matrix* \( G_i \), which is a *unit lower triangular matrix*, having ones on the diagonal and suitable possibly nonzero elements in rows \( i + 1, \ldots, n \) of column \( i \). This gives \( B^{(i)} = (G_iP_i)B^{(i-1)} \). Hence, we get, after some \( r \leq (n - 1) \) such operations,

\[
(G_rP_r) \cdots (G_2P_2)(G_1P_1)B = U. \tag{A.1}
\]

The system \( Bx = b \) can now be solved by computing \( \tilde{b} = (G_rP_r) \cdots (G_1P_1)b \) and then solving the triangular system \( Ux = \tilde{b} \) by back-substitution. If no permutations are performed, \( G_r \cdots G_1 \) is lower triangular, and denoting its (lower triangular) inverse as \( L \), we have the factored form \( B = LU \) for \( B \), hence its name. Also, if \( P' \) is a permutation matrix that is used to *a priori* rearrange the rows of \( B \) and we then apply the Gaussian triangularization operation to derive \( L^{-1}P'B = U \), we can write \( B = (P')^{-1}LU = PLU \), noting that \( P' = P^{-1} \). Hence, this factorization is sometimes called a *PLU decomposition*. If \( B \) is sparse, \( P' \) can be used to make \( P'B \) nearly upper triangular (assuming that the columns of \( B \) have been appropriately permuted) and then only a few and sparse Gaussian pivot operations will be required to obtain \( U \). This method is therefore very well suited for sparse matrices.

*QR and QRP Factorization for a Basis B*

This factorization is most suitable and is used frequently for solving dense equation systems. Here the matrix \( B \) is reduced to an upper triangular form \( R \) by premultiplying it with a sequence of square, symmetric orthogonal matrices \( Q_i \).

Given \( B^{(i-1)} = Q_{i-1} \cdots Q_1B \) that is upper triangular in columns \( 1, \ldots, i - 1 \) (where \( B^{(0)} = B \)), we construct a matrix \( Q_i \) so that \( Q_iB^{(i-1)} = B^{(i)} \) is upper triangular in column \( i \) as well, while columns \( 1, \ldots, i - 1 \) remain unaffected. The matrix \( Q_i \) is a square, symmetric orthogonal matrix of the form \( Q_i = I - \gamma_iq_iq_i' \), where \( q_i = (0, \ldots, 0, q_{ii}, \ldots, q_{ni})' \) and \( \gamma_i \in \mathbb{R} \) are suitably chosen to perform the foregoing operation. Such a matrix \( Q_i \) is called a *Householder transformation matrix*. If the elements in rows \( i, \ldots, n \) of column \( i \) of \( B^{(i-1)} \) are denoted by \((\alpha_i, \ldots, \alpha_n)'\), then we have \( q_{ii} = \alpha_i + \theta_i, q_{jj} = \alpha_j \) for \( j = i + 1, \ldots, n \), \( \gamma_i = 1/\theta_iq_{ii} \), where \( \theta_i = \text{sign}(\alpha_i)[\alpha_i^2 + \cdots + \alpha_n^2]^{1/2} \), and where \( \text{sign}(\alpha_i) = 1 \) if \( \alpha_i > 0 \) and \(-1 \) otherwise. Defining \( Q = Q_{n-1} \cdots Q_1 \), we see that \( Q \) is also a symmetric orthogonal matrix and that \( QB = R \), or that \( B = QR \), since \( Q = Q' = Q^{-1} \); that is, \( Q \) is an *involutory matrix*. 


Now, to solve $Bx = b$, we equivalently solve $QRx = b$ or $Rx = Qb$ by finding $\tilde{b} = Qb$ first and then solving the upper triangular system $Rx = \tilde{b}$ via back-substitution. Note that since $\|Qv\| = \|v\|$ for any vector $v$, we have $\|R\| = \|QR\| = \|B\|$, so that $R$ preserves the relative magnitudes of the elements in $B$, maintaining stability. This is its principal advantage.

Also, a permutation matrix $P$ is sometimes used to postmultiply $B(i-1)$ before applying $Q_i$ to it, so as to move a column that has the largest value of the sum of squares below row $i - 1$ into the $i$th column position (see the computation of $\theta_i$ above). Since the product of permutation matrices is also a permutation matrix, and since a permutation matrix is orthogonal, this leads to the decomposition $B = QRP$ via the operation sequence $Q_{n-1} \cdots Q_1BP_1P_2 \cdots P_{n-1} = R$.

### Cholesky Factorization $LL^T$ and $LDL^T$ for Symmetric, Positive Definite Matrices $B$

The Cholesky factorization of a symmetric, positive definite matrix $B$ represents this matrix as $B = LL^T$, where $L$ is a lower triangular matrix of the form

$$L = \begin{bmatrix} \ell_{11} & \ell_{21} & \ell_{22} & 0 \\ \ell_{31} & \ell_{32} & \ell_{33} & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \ell_{nn} \end{bmatrix},$$

so that

$$LL^T = \begin{bmatrix} \ell_{11}^2 & \ell_{21}\ell_{11} + \ell_{22}^2 \\ \ell_{31}\ell_{11} + \ell_{21}\ell_{31} + \ell_{22}\ell_{32} & \ell_{32}^2 + \ell_{33}^2 \\ \vdots & \vdots \\ \ell_{n1}\ell_{11} + \ell_{21}\ell_{n1} + \ell_{22}\ell_{n2} & \ell_{n2}^2 + \ell_{n3}^2 + \ell_{n4}^2 \\ \ell_{n1}\ell_{11} + \ell_{21}\ell_{n1} + \ell_{22}\ell_{n2} + \ell_{23}\ell_{n3} & \ell_{n2}^2 + \ell_{n3}^2 + \ell_{n4}^2 + \ell_{n5}^2 \\ \vdots & \vdots \\ \ell_{n1}\ell_{11} + \ell_{21}\ell_{n1} + \ell_{22}\ell_{n2} + \ell_{23}\ell_{n3} + \ell_{24}\ell_{n4} & \ell_{n2}^2 + \ell_{n3}^2 + \ell_{n4}^2 + \ell_{n5}^2 + \ell_{n6}^2 \\ \ell_{n1}\ell_{11} + \ell_{21}\ell_{n1} + \ell_{22}\ell_{n2} + \ell_{23}\ell_{n3} + \ell_{24}\ell_{n4} + \ell_{25}\ell_{n5} & \ell_{n2}^2 + \ell_{n3}^2 + \ell_{n4}^2 + \ell_{n5}^2 + \ell_{n6}^2 + \ell_{n7}^2 \\ \vdots & \vdots \\ \ell_{n1}\ell_{11} + \ell_{21}\ell_{n1} + \ell_{22}\ell_{n2} + \ell_{23}\ell_{n3} + \ell_{24}\ell_{n4} + \ell_{25}\ell_{n5} + \ell_{26}\ell_{n6} & \ell_{n2}^2 + \ell_{n3}^2 + \ell_{n4}^2 + \ell_{n5}^2 + \ell_{n6}^2 + \ell_{n7}^2 + \ell_{n8}^2 \\ \ell_{n1}\ell_{11} + \ell_{21}\ell_{n1} + \ell_{22}\ell_{n2} + \ell_{23}\ell_{n3} + \ell_{24}\ell_{n4} + \ell_{25}\ell_{n5} + \ell_{26}\ell_{n6} + \ell_{27}\ell_{n7} & \ell_{n2}^2 + \ell_{n3}^2 + \ell_{n4}^2 + \ell_{n5}^2 + \ell_{n6}^2 + \ell_{n7}^2 + \ell_{n8}^2 + \ell_{n9}^2 \end{bmatrix}$$

By equating the elements of $B$ directly to those in $LL^T$, we obtain the system of equations

$$\begin{align*}
\ell_{11}^2 &= b_{11}, & \ell_{21}\ell_{11} &= b_{21}, & \ell_{31}\ell_{11} &= b_{31}, & \cdots & \ell_{n1}\ell_{11} &= b_{n1} \\
\ell_{21}^2 + \ell_{22}^2 &= b_{22}, & \ell_{21}\ell_{31} + \ell_{22}\ell_{32} &= b_{32}, & \cdots & \ell_{21}\ell_{n1} + \ell_{22}\ell_{n2} &= b_{n2} \\
\ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 &= b_{33}, & \ell_{31}\ell_{m1} + \ell_{32}\ell_{m2} + \ell_{33}\ell_{m3} &= b_{m3} \\
\ell_{n1}^2 + \cdots + \ell_{nn}^2 &= b_{nn}.
\end{align*}$$

These equations can be used sequentially to compute the unknowns $\ell_{ij}$ in the order $\ell_{11}, \ell_{21}, \ldots, \ell_{nl}, \ell_{22}, \ell_{32}, \ldots, \ell_{n2}, \ell_{33}, \ldots, \ell_{nn}$, by using the equation
for $b_{ij}$ to compute $\ell_{ij}$ for $j = 1, \ldots, n$, $i = j, \ldots, n$. Note that these equations are well-defined for a symmetric, positive definite matrix $B$ and that $LL^t$ is positive definite if and only if $\ell_{ii} > 0$ for all $i = 1, \ldots, n$.

The equation system $Bx = b$ can now be solved via $L(L'x) = b$ through the solution of two triangular systems of equations. We first find $y$ to satisfy $Ly = b$ and then compute $x$ via the system $L'x = y$.

Sometimes the Cholesky factorization is represented as $B = LDL^t$, where $L$ is a lower triangular matrix (usually having ones along its diagonal) and $D$ is a diagonal matrix, both having positive diagonal entries. Writing $B = LDL^t = (LD^{1/2})(LD^{1/2})^t = L'L^t$, we see that the two representations are related equivalently. The advantage of the representation $LDL^t$ is that $D$ can be used to avoid the square root operation associated with the diagonal system of equations, and this improves the accuracy of computations. (For example, the diagonal components of $L$ can be made unity.)

Also, if $B$ is a general basis matrix, then since $BB^t$ is symmetric and positive definite, it has a Cholesky factorization $BB^t = LL^t$. In such a case, $L$ is referred to as the Cholesky factor associated with $B$. Note that we can determine $L$ in this case by finding the QR decomposition for $B^t$ so that $BB^t = R'Q'QR = R'R$, and therefore, $L = R^t$. Whenever this is done, note that the matrix $Q$ or its components $Q_i$ need not be stored, since we are only interested in the resulting upper triangular matrix $R$.

### A.3 Sets and Sequences

A set is a collection of elements or objects. A set may be specified by listing its elements or by specifying the properties that the elements must satisfy. For example, the set $S = \{1, 2, 3, 4\}$ can be represented alternatively as $S = \{x : 1 \leq x \leq 4, x \text{ integer}\}$. If $x$ is a member of $S$, we write $x \in S$, and if $x$ is not a member of $S$, we write $x \notin S$. Sets are denoted by capital letters, such as $S$, $X$, and $A$. The empty set, denoted by $\emptyset$, has no elements.

**Unions, Intersections, and Subsets**

Given two sets, $S_1$ and $S_2$, the set consisting of elements that belong to either $S_1$ or $S_2$, or both, is called the union of $S_1$ and $S_2$ and is denoted by $S_1 \cup S_2$. The elements belonging to both $S_1$ and $S_2$ form the intersection of $S_1$ and $S_2$, denoted $S_1 \cap S_2$. If $S_1$ is a subset of $S_2$, that is, if each element of $S_1$ is also an element of $S_2$, we write $S_1 \subseteq S_2$ or $S_2 \supseteq S_1$. Thus, we write $S \subseteq R^n$ to denote
that all elements in $S$ are points in $R^n$. A strict containment $S_1 \subset S_2$, $S_1 \neq S_2$, is denoted by $S_1 \subset S_2$.

**Closed and Open Intervals**

Let $a$ and $b$ be two real numbers. The closed interval $[a, b]$ denotes all real numbers satisfying $a \leq x \leq b$. Real numbers satisfying $a \leq x < b$ are represented by $[a, b)$, while those satisfying $a < x \leq b$ are denoted by $(a, b)$. Finally, the set of points $x$ with $a < x < b$ is represented by the open interval $(a, b)$.

**Greatest Lower Bound and Least Upper Bound**

Let $S$ be a set of real numbers. Then the greatest lower bound, or the infimum, of $S$ is the largest possible scalar $\alpha$ satisfying $\alpha \leq x$ for each $x \in S$. The infimum is denoted by $\inf \{x : x \in S\}$. The least upper bound, or the supremum, of $S$ is the smallest possible scalar $\alpha$ satisfying $\alpha \geq x$ for each $x \in S$. The supremum is denoted by $\sup \{x : x \in S\}$.

**Neighborhoods**

Given a point $x \in R^n$ and an $\varepsilon > 0$, the ball $N_{\varepsilon}(x) = \{y : \|y - x\| \leq \varepsilon\}$ is called an $\varepsilon$-neighborhood of $x$. The inequality in the definition of $N_{\varepsilon}(x)$ is sometimes replaced by a strict inequality.

**Interior Points and Open Sets**

Let $S$ be a subset of $R^n$, and let $x \in S$. Then $x$ is called an interior point of $S$ if there is an $\varepsilon$-neighborhood of $x$ that is contained in $S$, that is, if there exists an $\varepsilon > 0$ such that $\|y - x\| \leq \varepsilon$ implies that $y \in S$. The set of all such points is called the interior of $S$ and is denoted by $\text{int } S$. Furthermore, $S$ is called open if $S = \text{int } S$.

**Relative Interior**

Let $S \subset R^n$, and let $\text{aff}(S)$ denote the affine hull of $S$. Although $\text{int}(S) = \emptyset$, the interior of $S$ as viewed in the space of its affine hull may be nonempty. This is called the relative interior of $S$ and is denoted by $\text{relint}(S)$. Specifically, $\text{relint}(S) = \{x \in \text{aff}(S) : \text{aff}(S) \subset S \text{ for some } \varepsilon > 0\}$. Note that if $S_1 \subseteq S_2$, $\text{relint}(S_1)$ is not necessarily contained within $\text{relint}(S_2)$, although $\text{int}(S_1) \subseteq \text{int}(S_2)$. For example, if $S_1 = \{x : \alpha^T x = \beta\}$, $\alpha \neq 0$ and $S_2 = \{x : \alpha^T x \leq \beta\}$, $S_1 \subseteq S_2$, $\text{int}(S_1) = \emptyset \subseteq \text{int}(S_2) = \{x : \alpha^T x < \beta\}$, but $\text{relint}(S_1) = S_1 \not\subseteq \text{relint}(S_2) = \text{int}(S_2)$. 
Bounded Sets

A set \( S \subset \mathbb{R}^n \) is said to be **bounded** if it can be contained within a ball of finite radius.

Closure Points and Closed Sets

Let \( S \) be a subset of \( \mathbb{R}^n \). The **closure** of \( S \), denoted \( \text{cl} \ S \), is the set of all points that are arbitrarily close to \( S \). In particular, \( x \in \text{cl} \ S \) if for each \( \varepsilon > 0 \), \( S \cap N_\varepsilon(x) \neq \emptyset \), where \( N_\varepsilon(x) = \{ y : \| y - x \| \leq \varepsilon \} \). The set \( S \) is said to be **closed** if \( S = \text{cl} \ S \).

Boundary Points

Let \( S \) be a subset of \( \mathbb{R}^n \). Then \( x \) is called a **boundary point** of \( S \) if for each \( \varepsilon > 0 \), \( N_\varepsilon(x) \) contains a point in \( S \) and a point not in \( S \), where \( N_\varepsilon(x) = \{ y : \| y - x \| \leq \varepsilon \} \). The set of all boundary points is called the **boundary** of \( S \) and is denoted by \( \partial S \).

Sequences and Subsequences

A **sequence** of vectors \( x_1, x_2, x_3, \ldots \), is said to **converge** to the **limit point** \( \bar{x} \) if \( \| x_k - \bar{x} \| \to 0 \) as \( k \to \infty \); that is, if for any given \( \varepsilon > 0 \), there is a positive integer \( N \) such that \( \| x_k - \bar{x} \| < \varepsilon \) for all \( k \geq N \). The sequence is usually denoted by \( \{ x_k \} \), and the limit point \( \bar{x} \) is represented by either \( x_k \to \bar{x} \) as \( k \to \infty \) or by \( \lim_{k \to \infty} x_k = \bar{x} \). Any converging sequence has a unique limit point.

By deleting certain elements of a sequence \( \{ x_k \} \), we obtain a **subsequence**. A subsequence is usually denoted as \( \{ x_k \}_{\mathcal{K}} \) where \( \mathcal{K} \) is a subset of all positive integers. To illustrate, let \( \mathcal{K} \) be the set of all even positive integers. Then \( \{ x_k \}_{\mathcal{K}} \) denotes the subsequence \( \{ x_2, x_4, x_6, \ldots \} \).

Given a subsequence \( \{ x_k \}_{\mathcal{K}} \), the notation \( \{ x_{k+1} \}_{\mathcal{K}} \) denotes the subsequence obtained by adding 1 to the indices of all elements in the subsequence \( \{ x_k \}_{\mathcal{K}} \). To illustrate, if \( \mathcal{K} = \{ 3, 5, 10, 15, \ldots \} \), then \( \{ x_{k+1} \}_{\mathcal{K}} \) denotes the subsequence \( \{ x_4, x_6, x_{11}, x_{16}, \ldots \} \).

A sequence \( \{ x_k \} \) is called a **Cauchy sequence** if for any given \( \varepsilon > 0 \), there is a positive integer \( N \) such that \( \| x_k - x_m \| < \varepsilon \) for all \( k, m \geq N \). A sequence in \( \mathbb{R}^n \) has a limit if and only if it is Cauchy.

Let \( \{ x_n \} \) be a bounded sequence in \( \mathbb{R} \). The **limit superior** of \( \{ x_n \} \), denoted \( \limsup(x_n) \) or \( \overline{\lim}(x_n) \), equals the infimum of all numbers \( q \in \mathbb{R} \) for which at most a finite number of the elements of \( \{ x_n \} \) (strictly) exceed \( q \). Similarly, the **limit inferior** of \( \{ x_n \} \) is given by \( \liminf(x_n) = \underline{\lim}(x_n) = \sup\{ q : \text{at most a finite} \)
number of elements of \( \{x_n\} \) are (strictly) less than \( q \). A bounded sequence always has a unique \( \limsup \) and \( \liminf \).

Compact Sets

A set \( S \) in \( \mathbb{R}^n \) is said to be \textit{compact} if it is closed and bounded. For every sequence \( \{x_k\} \) in a compact set \( S \), there is a convergent subsequence with a limit in \( S \).

A.4 Functions

A \textit{real-valued function} \( f \) defined on a subset \( S \) of \( \mathbb{R}^n \) associates with each point \( x \) in \( S \) a real number \( f(x) \). The notation \( f: S \to \mathbb{R} \) denotes that the domain of \( f \) is \( S \) and that the range is a subset of the real numbers. If \( f \) is defined everywhere on \( \mathbb{R}^n \) or if the domain is not important, the notation \( f: \mathbb{R}^n \to \mathbb{R} \) is used. A collection of real-valued functions \( f_1, \ldots, f_m \) can be viewed as a single \textit{vector function} \( f \) whose \( j \)th component is \( f_j \).

Continuous Functions

A function \( f: S \to \mathbb{R} \) is said to be \textit{continuous at} \( \bar{x} \in S \) if for any given \( \varepsilon > 0 \), there is a \( \delta > 0 \) such that \( x \in S \) and \( \|x - \bar{x}\| < \delta \) imply that \( |f(x) - f(\bar{x})| < \varepsilon \). Equivalently, \( f \) is continuous at \( \bar{x} \in S \), if for any sequence \( \{x_n\} \to \bar{x} \) such that \( \{f(x_n)\} \to \bar{f} \), we have that \( f(\bar{x}) = \bar{f} \) as well. A vector-valued function is said to be continuous at \( \bar{x} \) if each of its components is continuous at \( \bar{x} \).

Upper and Lower Semicontinuity

Let \( S \) be a nonempty set in \( \mathbb{R}^n \). A function \( f: S \to \mathbb{R} \) is said to be \textit{upper semicontinuous at} \( \bar{x} \in S \) if for each \( \varepsilon > 0 \) there exists a \( \delta > 0 \) such that \( x \in S \) and \( \|x - \bar{x}\| < \delta \) imply that \( f(x) - f(\bar{x}) < \varepsilon \). Similarly, a function \( f: \mathbb{R}^n \to \mathbb{R} \) is called \textit{lower semicontinuous at} \( \bar{x} \in S \) if for each \( \varepsilon > 0 \) there exists a \( \delta > 0 \) such that \( x \in S \) and \( \|x - \bar{x}\| < \delta \) imply that \( f(x) - f(\bar{x}) > -\varepsilon \). Equivalently, then \( f \) is \textit{upper semicontinuous at} \( \bar{x} \in S \), if, for any sequence \( \{x_n\} \to \bar{x} \) such that \( \{f(x_n)\} \to \bar{f} \), we have that \( f(\bar{x}) \geq \bar{f} \). Similarly, if \( f(\bar{x}) \leq \bar{f} \) for any such sequence, then \( f \) is said to be \textit{lower semicontinuous at} \( \bar{x} \). Hence, a function is \textit{continuous at} \( \bar{x} \) if and only if it is both upper and lower semicontinuous at \( \bar{x} \). A vector-valued function is called upper or lower semicontinuous if each of its components is upper or lower semicontinuous, respectively.

Minima and Maxima of Semicontinuous Functions

Let \( S \) be a nonempty compact set in \( \mathbb{R}^n \) and suppose that \( f: \mathbb{R}^n \to \mathbb{R} \). If \( f \) is lower semicontinuous, then it assumes a minimum over \( S \); that is, there exists an \( \bar{x} \in S \)
such that \( f(x) \leq f(x) \) for each \( x \in S \). Similarly, if \( f \) is upper semicontinuous, then it assumes a maximum over \( S \). Since a continuous function is both lower and upper semicontinuous, it achieves both a minimum and a maximum over any nonempty compact set.

**Differentiable Functions**

Let \( S \) be a nonempty set in \( \mathbb{R}^n \), \( \bar{x} \in \text{int} S \) and let \( f: S \to \mathbb{R} \). Then \( f \) is said to be **differentiable at** \( \bar{x} \) if there is a vector \( \nabla f(\bar{x}) \) in \( \mathbb{R}^n \) called the **gradient** of \( f \) at \( \bar{x} \) and a function \( \beta \) satisfying \( \beta(\bar{x}; x) \to 0 \) as \( x \to \bar{x} \) such that

\[
f(x) = f(\bar{x}) + \nabla f(\bar{x})' (x - \bar{x}) + \|x - \bar{x}\| \beta(\bar{x}; x) \quad \text{for each } x \in S.
\]

The gradient vector consists of the partial derivatives, that is,

\[
\nabla f(\bar{x})' = \left( \frac{\partial f(\bar{x})}{\partial x_1}, \frac{\partial f(\bar{x})}{\partial x_2}, \ldots, \frac{\partial f(\bar{x})}{\partial x_n} \right).
\]

Furthermore, \( f \) is called **twice differentiable at** \( \bar{x} \) if, in addition to the gradient vector, there exist an \( n \times n \) symmetric matrix \( H(\bar{x}) \), called the **Hessian matrix** of \( f \) at \( \bar{x} \), and a function \( \beta \) satisfying \( \beta(\bar{x}; x) \to 0 \) as \( x \to \bar{x} \) such that

\[
f(x) = f(\bar{x}) + \nabla f(\bar{x})' (x - \bar{x}) + \frac{1}{2} (x - \bar{x})' H(\bar{x})(x - \bar{x}) + \|x - \bar{x}\|^2 \beta(\bar{x}; x)
\]

for each \( x \in S \).

The element in row \( i \) and column \( j \) of the Hessian matrix is the second partial \( \frac{\partial^2 f(\bar{x})}{\partial x_i \partial x_j} \).

A vector-valued function is differentiable if each of its components is differentiable and is twice differentiable if each of its components is twice differentiable.

In particular, for a differentiable vector function \( h: \mathbb{R}^n \to \mathbb{R}^\ell \) where \( h(x) = (h_1(x), \ldots, h_\ell(x))^t \), the **Jacobian** of \( h \), denoted by the gradient notation \( \nabla h(x) \), is given by the \( \ell \times n \) matrix

\[
\nabla h(x) = \begin{bmatrix}
\nabla h_1(x)^t \\
\vdots \\
\nabla h_\ell(x)^t
\end{bmatrix}_{\ell \times n},
\]

whose rows correspond to the transpose of the gradients of \( h_1, \ldots, h_\ell \), respectively.
Mean Value Theorem

Let $S$ be a nonempty open convex set in $\mathbb{R}^n$, and let $f: S \to \mathbb{R}$ be differentiable. The mean value theorem can be stated as follows. For every $x_1$ and $x_2$ in $S$, we must have

$$f(x_2) = f(x_1) + \nabla f(x)^T (x_2 - x_1),$$

where $x = \lambda x_1 + (1 - \lambda)x_2$ for some $\lambda \in (0, 1)$.

Taylor's Theorem

Let $S$ be a nonempty open convex set in $\mathbb{R}^n$, and let $f: S \to \mathbb{R}$ be twice differentiable. The second-order form of Taylor's theorem can be stated as follows. For every $x_1$ and $x_2$ in $S$, we must have

$$f(x_2) = f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + \frac{1}{2} (x_2 - x_1)^T H(x)(x_2 - x_1),$$

where $H(x)$ is the Hessian of $f$ at $x$, and where $x = \lambda x_1 + (1 - \lambda)x_2$ for some $\lambda \in (0, 1)$.
Appendix B

Summary of Convexity, Optimality Conditions, and Duality

This appendix gives a summary of the relevant results from Chapters 2 through 6 on convexity, optimality conditions, and duality. It is intended to provide the minimal background needed for an adequate coverage of Chapters 8 through 11, excluding convergence analysis.

B.1 Convex Sets

A set $S$ in $\mathbb{R}^n$ is said to be convex if for each $x_1, x_2 \in S$, the line segment $\lambda x_1 + (1-\lambda)x_2$ for $\lambda \in [0, 1]$ belongs to $S$. Points of the form $x = \lambda x_1 + (1-\lambda)x_2$ for $\lambda \in [0, 1]$ are called convex combinations of $x_1$ and $x_2$. Figure B.1 illustrates an example of a convex set and an example of a nonconvex set.

We present below some examples of convex sets frequently encountered in mathematical programming.

1. **Hyperplane**: $S = \{x : p'x = \alpha\}$, where $p$ is a nonzero vector in $\mathbb{R}^n$, called the normal to the hyperplane, and $\alpha$ is a scalar.
2. **Half-space**: $S = \{x : p'x \leq \alpha\}$, where $p$ is a nonzero vector in $\mathbb{R}^n$ and $\alpha$ is a scalar.
3. **Open half-space**: $S = \{x : p'x < \alpha\}$, where $p$ is a nonzero vector in $\mathbb{R}^n$ and $\alpha$ is a scalar.
4. **Polyhedral set**: $S = \{x : Ax \leq b\}$, where $A$ is an $m \times n$ matrix and $b$ is an $m$-vector.
5. **Polyhedral cone**: $S = \{x : Ax \leq 0\}$, where $A$ is an $m \times n$ matrix.
6. **Cone spanned by a finite number of vectors**: $S = \{x : x = \sum_{j=1}^{m} \lambda_j a_j, \lambda_j \geq 0 \text{ for } j = 1, \ldots, m\}$, where $a_1, \ldots, a_m$ are given vectors in $\mathbb{R}^n$.
7. **Neighborhood**: $S = \{x : \|x - \bar{x}\| \leq \varepsilon\}$, where $\bar{x}$ is a fixed vector in $\mathbb{R}^n$ and $\varepsilon > 0$.  

765
Given two nonempty convex sets \( S_1 \) and \( S_2 \) in \( \mathbb{R}^n \) such that \( S_1 \cap S_2 = \emptyset \), there exists a hyperplane \( H = \{ x : p^T x = \alpha \} \) that separates them; that is,

\[
p^T x \leq \alpha \quad \text{for all } x \in S_1 \quad \text{and} \quad p^T x \geq \alpha \quad \text{for all } x \in S_2.
\]

Here \( H \) is called a separating hyperplane whose normal is the nonzero vector \( p \).

Closely related to the above concept is the notion of a supporting hyperplane. Let \( S \) be a nonempty convex set in \( \mathbb{R}^n \), and let \( \bar{x} \) be a boundary point. Then there exists a hyperplane \( H = \{ x : p^T x = \alpha \} \) that supports \( S \) at \( \bar{x} \); that is,

\[
p^T \bar{x} = \alpha \quad \text{and} \quad p^T x \leq \alpha \quad \text{for all } x \in S.
\]

In Figure B.2 we illustrate the concepts of separating and supporting hyperplanes.

The following two theorems are used in proving optimality conditions and duality relationships and in developing termination criteria for algorithms.
Farkas’s Theorem

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

\begin{align*}
\text{System 1} & \quad Ax \leq 0, \ c'x > 0 \quad \text{for some } x \in \mathbb{R}^n. \\
\text{System 2} & \quad A'y = c, \ y \geq 0 \quad \text{for some } y \in \mathbb{R}^m.
\end{align*}

Gordan’s Theorem

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

\begin{align*}
\text{System 1} & \quad Ax < 0 \quad \text{for some } x \in \mathbb{R}^n. \\
\text{System 2} & \quad A'y = 0, \ y \geq 0 \quad \text{for some nonzero } y \in \mathbb{R}^m.
\end{align*}

An important concept in convexity is that of an extreme point. Let $S$ be a non-empty convex set in $\mathbb{R}^n$. A vector $x \in S$ is called an extreme point of $S$ if $x = \lambda x_1 + (1 - \lambda)x_2$ with $x_1, x_2 \in S$, and $\lambda \in (0, 1)$ implies that $x = x_1 = x_2$. In other words, $x$ is an extreme point if it cannot be represented as a strict convex combination of two distinct points in $S$. In particular, for the set $S = \{x : Ax = b, x \geq 0\}$, where $A$ is an $m \times n$ matrix of rank $m$ and $b$ is an $m$-vector, $x$ is an extreme point of $S$ if and only if the following conditions hold true. The matrix $A$ can be decomposed into $[B, N]$, where $B$ is an $m \times m$ invertible matrix and $x^I = (x_B^I, x_N^I)$, where $x_B = B^{-1}b \geq 0$ and $x_N = 0$.

Another concept that is used in the case of an unbounded convex set is that of a direction of the set. Specifically, if $S$ is an unbounded closed convex set, a vector $d$ is a direction of $S$ if $x + \lambda d \in S$ for each $\lambda \geq 0$ and for each $x \in S$.

B.2 Convex Functions and Extensions

Let $S$ be a nonempty convex set in $\mathbb{R}^n$. The function $f: S \to \mathbb{R}$ is said to be convex on $S$ if

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

for each $x_1, x_2 \in S$ and for each $\lambda \in [0, 1]$. The function $f$ is said to be strictly convex on $S$ if the above inequality holds as a strict inequality for each distinct $x_1, x_2 \in S$ and for each $\lambda \in (0, 1)$. The function $f$ is said to be concave (strictly concave) if $-f$ is convex (strictly convex). Figure B.3 shows some examples of convex and concave functions.

Following are some examples of convex functions. By taking the negatives of these functions, we get some examples of concave functions.
In many cases, the assumption of convexity of a function can be relaxed to the weaker notions of quasiconvex and pseudoconvex functions.

Let $S$ be a nonempty convex set in $\mathbb{R}^n$. The function $f: S \to \mathbb{R}$ is said to be **quasiconvex** on $S$ if for each $x_1, x_2 \in S$, the following inequality holds true:

$$f[\lambda x_1 + (1-\lambda)x_2] \leq \max\{f(x_1), f(x_2)\} \quad \text{for each } \lambda \in (0, 1).$$

The function $f$ is said to be **strictly quasiconvex** on $S$ if the above inequality holds as a strict inequality, provided that $f(x_1) \neq f(x_2)$. The function $f$ is said to be **strongly quasiconvex** on $S$ if the above inequality holds as a strict inequality for $x_1 \neq x_2$.

Let $S$ be a nonempty open convex set in $\mathbb{R}^n$. The function $f: S \to \mathbb{R}$ is said to be **pseudoconvex** if for each $x_1, x_2 \in S$ with $\nabla f(x_1)^T (x_2 - x_1) \geq 0$, we have $f(x_2) \geq f(x_1)$. The function $f$ is said to be **strictly pseudoconvex** on $S$ if whenever $x_1$ and $x_2$ are distinct points in $S$ with $\nabla f(x_1)^T (x_2 - x_1) \geq 0$, we have $f(x_2) > f(x_1)$.

The above generalizations of convexity extend to the concave case by replacing $f$ by $-f$. Figure B.4 illustrates these concepts. Figure B.5 summarizes the relationships among different types of convexity.

We now give a summary of important properties for various types of convex functions. Here $f: S \to \mathbb{R}$, where $S$ is a nonempty convex set in $\mathbb{R}^n$. 

---

**Figure B.3 Convex and concave functions.**

1. $f(x) = 3x + 4$.
2. $f(x) = |x|$.
3. $f(x) = x^2 - 2x$.
4. $f(x) = -x^{1/2}$ for $x \geq 0$.
5. $f(x_1, x_2) = 2x_1^2 + x_2^2 - 2x_1x_2$.
6. $f(x_1, x_2, x_3) = x_1^4 + 2x_2^2 + 3x_3^2 - 4x_1 - 4x_2x_3$. 

In many cases, the assumption of convexity of a function can be relaxed to the weaker notions of quasiconvex and pseudoconvex functions.
Summary of Convexity, Optimality Conditions and Duality

Both quasiconvex and pseudoconvex

Quasiconvex but not pseudoconvex

Neither quasiconvex nor pseudoconvex

Figure B.4 Quasiconvexity and pseudoconvexity.

Strictly Convex Functions

1. The function $f$ is continuous on the interior of $S$.
2. The set $\{(x, y) : x \in S, y \geq f(x)\}$ is convex.
3. The set $\{x \in S : f(x) \leq \alpha\}$ is convex for each real $\alpha$.
4. A differentiable function $f$ is strictly convex on $S$ if and only if $f(x) > f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x})$ for each distinct $x, \bar{x} \in S$.
5. Let $f$ be twice differentiable. Then if the Hessian $H(x)$ is positive definite for each $x \in S$, $f$ is strictly convex on $S$.
6. Every local minimum of $f$ over a convex set $X \subseteq S$ is the unique global minimum.
7. If $\nabla f(\bar{x}) = 0$, then $\bar{x}$ is the unique global minimum of $f$ over $S$.
8. The maximum of $f$ over a nonempty compact polyhedral set $X \subseteq S$ is achieved at an extreme point of $X$.

Convex Functions

1. The function $f$ is continuous on the interior of $S$.
2. The function $f$ is convex on $S$ if and only if the set $\{(x, y) : x \in S, y \geq f(x)\}$ is convex.
3. The set $\{x \in S : f(x) \leq \alpha\}$ is convex for each real $\alpha$.
4. A differentiable function $f$ is convex on $S$ if and only if $f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x})$ for each $x, \bar{x} \in S$.
5. A twice differentiable function $f$ is convex on $S$ if and only if the Hessian $H(x)$ is positive semidefinite for each $x \in S$.
6. Every local minimum of $f$ over a convex set $X \subseteq S$ is a global minimum.
7. If $\nabla f(\bar{x}) = 0$, then $\bar{x}$ is a global minimum of $f$ over $S$.
Appendix B

Figure B.5 Relationship among various types of convexity.

8. A maximum of \( f \) over a nonempty compact polyhedral set \( X \subseteq S \) is achieved at an extreme point of \( X \).

Pseudoconvex Functions

1. The set \( \{ x \in S : f(x) \leq \alpha \} \) is convex for each real \( \alpha \).
2. Every local minimum of \( f \) over a convex set \( X \subseteq S \) is a global minimum.
3. If \( \nabla f(\bar{x}) = 0 \), then \( \bar{x} \) is a global minimum of \( f \) over \( S \).
4. A maximum of \( f \) over a nonempty compact polyhedral set \( X \subseteq S \) is achieved at an extreme point of \( X \).
5. This characterization and the next relate to twice differentiable functions \( f \) defined on an open convex set \( S \subseteq \mathbb{R}^n \), with Hessian \( H(x) \).

The function \( f \) is pseudoconvex on \( S \) if \( H(x) + r(x)\nabla f(x)\nabla f(x)' \) is positive semidefinite for all \( x \in S \), where \( r(x) = (1/2)[\delta - f(x)] \) for some \( \delta > f(x) \). Moreover, this condition is both necessary and sufficient if \( f \) is quadratic.

6. Define the \((n+1) \times (n+1)\) bordered Hessian \( B(x) \) of \( f \) as follows, where \( H(x) \) is "bordered" by an additional row and column:
Given any \( k \in \{1, \ldots, n\} \), and \( \gamma = \{i_1, \ldots, i_k\} \) composed of some \( k \) distinct indices \( 1 < i_1 < i_2 < \cdots < i_k \leq n \), the \textit{principal submatrix} \( B_{\gamma,k}(x) \) is a \((k+1) \times (k+1)\) submatrix of \( B(x) \) formed by picking the elements of \( B(x) \) that intersect in the rows \( i_1, \ldots, i_k, (n+1) \) and the columns \( i_1, \ldots, i_k, (n+1) \) of \( B(x) \). The \textit{leading principal submatrix} of \( B(x) \) is denoted by \( B_k(x) \) and equals \( B_{\gamma,k} \) for \( \gamma = \{1, \ldots, k\} \). Similarly, let \( H_{\gamma,k}(x) \) and \( H_k(x) \) be the \( k \times k \) principal submatrix and the leading principal submatrix, respectively, of \( H(x) \).

Then \( f \) is pseudoconvex on \( S \) if for each \( x \in S \), we have (i) \( \det B_{\gamma,k}(x) \leq 0 \) for all \( \gamma, k = 1, \ldots, n \), and (ii) if \( \det B_{\gamma,k}(x) = 0 \) for any \( \gamma, k \), then \( \det H_{\gamma,k} \geq 0 \) over some neighborhood of \( x \). Moreover, if \( f \) is quadratic, then these conditions are both necessary and sufficient. Also, in general, the condition \( \det B_k(x) \leq 0 \) for all \( k = 1, \ldots, n, x \in S \), is sufficient for \( f \) to be pseudoconvex on \( S \).

7. Let \( f: S \subseteq \mathbb{R}^n \rightarrow \mathbb{R} \) be quadratic, where \( S \) is a convex subset of \( \mathbb{R}^n \). Then \([f \text{ is pseudoconvex on } S] \Leftrightarrow \) [the bordered Hessian \( B(x) \) has exactly one simple negative eigenvalue for all \( x \in S \) \( \Leftrightarrow \) [for each \( y \in \mathbb{R}^n \) such that \( \nabla f(x)^T y = 0 \), we have that \( y^T H(x) y \geq 0 \) for all \( x \in S \)]. Moreover, \([f \text{ is strictly pseudoconvex on } S] \Leftrightarrow \) [for all \( x \in S \), and for all \( k = 1, \ldots, n \), we have (i) \( \det B_k(x) \leq 0 \), and (ii) if \( \det B_k(x) = 0 \), then \( \det H_k > 0 \)].

\textbf{Quasiconvex Functions}

1. The function \( f \) is quasiconvex over \( S \) if and only if \( \{x \in S : f(x) \leq \alpha\} \) is convex for each real \( \alpha \).
2. A maximum of \( f \) over a nonempty compact polyhedral set \( X \subseteq S \) is achieved at an extreme point of \( X \).
3. A differentiable function \( f \) on \( S \) is quasiconvex over \( S \) if and only if \( x_1, x_2 \in S \) with \( f(x_1) \leq f(x_2) \) implies that \( \nabla f(x_2)^T (x_1 - x_2) \leq 0 \).
4. Let \( f: S \subseteq \mathbb{R}^n \rightarrow \mathbb{R} \), where \( f \) is twice differentiable and \( S \) is a \textit{solid} (i.e., has a nonempty interior) convex subset of \( \mathbb{R}^n \). Define the bordered Hessian of \( f \) and its submatrices as in Property 6 of pseudoconvex functions. Then a sufficient condition for \( f \) to be quasiconvex on \( S \) is that for each \( x \in S \), \( \det B_k(x) < 0 \) for all \( k = 1, \ldots, n \). (Note that this condition actually implies that \( f \) is pseudoconvex.)
On the other hand, a necessary condition for \( f \) to be quasiconvex on \( S \) is that for each \( x \in S \), \( \det B_k(x) \leq 0 \) for all \( k = 1, \ldots, n \).

5. Let \( f : S \subseteq \mathbb{R}^n \to \mathbb{R} \) be a quadratic function where \( S \subseteq \mathbb{R}^n \) is a solid (nonempty interior) convex subset of \( \mathbb{R}^n \). Then \( f \) is quasiconvex on \( S \) if and only if \( f \) is pseudoconvex on \( \text{int}(S) \).

A local minimum of a strictly quasiconvex function over a convex set \( X \subseteq S \) is also a global minimum. Furthermore, if the function is strongly quasiconvex, the minimum is unique. If a function \( f \) is both strictly quasiconvex and lower semicontinuous, it is quasiconvex, so that the above properties for quasiconvexity hold true.

### B.3 Optimality Conditions

Consider the following problem:

\[
P: \quad \text{Minimize } f(x) \\
\text{subject to } g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m \\
\quad h_i(x) = 0 \quad \text{for } i = 1, \ldots, \ell \\
\quad x \in X,
\]

where \( f, g_i, h_i : \mathbb{R}^n \to \mathbb{R} \) and \( X \) is a nonempty open set in \( \mathbb{R}^n \). We give below the Fritz John necessary optimality conditions. If a point \( \bar{x} \) is a local optimal solution to the above problem, then there must exist a nonzero vector \((u_0, u, v)\) such that

\[
\begin{align*}
    u_0 \nabla f(\bar{x}) + \sum_{i=1}^{m} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^{\ell} v_i \nabla h_i(\bar{x}) &= 0 \\
    u_i g_i(\bar{x}) &= 0 \quad \text{for } i = 1, \ldots, m \\
    u_0 &\geq 0, u_i \geq 0 \quad \text{for } i = 1, \ldots, m,
\end{align*}
\]

where \( u \) and \( v \) are \( m \)- and \( \ell \)-vectors whose \( i \)th components are \( u_i \) and \( v_i \), respectively. Here, \( u_0, u_i, \) and \( v_i \) are referred to as the Lagrange or Lagrangian multipliers associated, respectively, with the objective function, the \( i \)th inequality constraint \( g_i(x) \leq 0 \), and the \( i \)th equality constraint \( h_i(x) = 0 \). The condition \( u_i g_i(\bar{x}) = 0 \) is called the complementary slackness condition and stipulates that either \( u_i = 0 \) or \( g_i(\bar{x}) = 0 \). Thus, if \( g_i(\bar{x}) < 0 \), then \( u_i = 0 \). By letting \( I \) be the set of binding inequality constraints at \( \bar{x} \), that is, \( I = \{i : g_i(\bar{x}) = 0\} \), then the Fritz John conditions can be written in the following equivalent form. If \( \bar{x} \) is a local optimal solution to Problem \( P \) above, then there exists a nonzero vector \((u_0, u_I, v)\) satisfying the following, where \( u_I \) is the vector of Lagrange multipliers associated with \( g_i(x) \leq 0 \) for \( i \in I \):
\[
\begin{align*}
&u_0 \nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^\ell v_i \nabla h_i(\bar{x}) = 0 \\
u_0 \geq 0, &\quad u_i \geq 0 \quad \text{for } i \in I.
\end{align*}
\]

If \( u_0 = 0 \), the Fritz John conditions become less meaningful, since essentially, they simply state that the gradients of the binding inequality constraints and the gradients of the equality constraints are linearly dependent. Under suitable assumptions, referred to as constraint qualifications, \( u_0 \) is guaranteed to be positive, and the Fritz John conditions reduce to the Karush–Kuhn–Tucker (KKT) conditions. A typical constraint qualification is that the gradients of the inequality constraints for \( i \in I \) and the gradients of the equality constraints at \( \bar{x} \) are linearly independent.

The KKT necessary optimality conditions can be stated as follows. If \( \bar{x} \) is a local optimal solution to Problem P, under a suitable constraint qualification, there exists a vector \((u, v)\) such that

\[
\begin{align*}
&\nabla f(\bar{x}) + \sum_{i=1}^m u_i \nabla g_i(\bar{x}) + \sum_{i=1}^\ell v_i \nabla h_i(\bar{x}) = 0 \\
u_1 g_1(\bar{x}) = 0 &\quad \text{for } i = 1, \ldots, m \\
u_i \geq 0 &\quad \text{for } i = 1, \ldots, m.
\end{align*}
\]

Again, \( u_i \) and \( v_i \) are the Lagrange or Lagrangian multipliers associated with the constraints \( g_i(x) \leq 0 \) and \( h_i(x) = 0 \), respectively. Furthermore, \( u_i g_i(\bar{x}) = 0 \) is referred to as a complementary slackness condition. If we let \( I = \{i : g_i(\bar{x}) = 0\} \), the above conditions can be rewritten as

\[
\begin{align*}
&\nabla f(\bar{x}) + \sum_{i=1}^m u_i \nabla g_i(\bar{x}) + \sum_{i=1}^\ell v_i \nabla h_i(\bar{x}) = 0 \\
u_i \geq 0 &\quad \text{for } i \in I.
\end{align*}
\]

Under suitable convexity assumptions, the KKT conditions are also sufficient for optimality. In particular, suppose that \( \bar{x} \) is a feasible solution to Problem P and that the KKT conditions stated below hold true:

\[
\begin{align*}
&\nabla f(\bar{x}) + \sum_{i \in I} u_i \nabla g_i(\bar{x}) + \sum_{i=1}^\ell v_i \nabla h_i(\bar{x}) = 0 \\
u_i \geq 0 &\quad \text{for } i \in I,
\end{align*}
\]

where \( I = \{i : g_i(\bar{x}) = 0\} \). If \( f \) is pseudoconvex, \( g_i \) is quasiconvex for \( i \in I \); and if \( h_i \) is quasiconvex if \( v_i > 0 \) and quasiconcave if \( v_i < 0 \), then \( \bar{x} \) is an optimal solution to Problem P.

To illustrate the KKT conditions, consider the following problem:
Minimize \((x_1 - 3)^2 + (x_2 - 2)^2\)
subject to
\[x_1^2 + x_2^2 \leq 5\]
\[x_1 + 2x_2 \leq 4\]
\[-x_1 \leq 0\]
\[-x_2 \leq 0.\]

The problem is illustrated in Figure B.6. Note that the optimal solution is \(\bar{x} = (2, 1)^t\). We first verify that the KKT conditions hold true at \(\bar{x}\). Here, the set of binding inequality constraints is \(I = \{1, 2\}\), so that we must have \(u_3 = u_4 = 0\) to satisfy the complementary slackness conditions. Note that

\[
\nabla f(\bar{x}) = (-2, -2)^t, \quad \nabla g_1(\bar{x}) = (4, 2)^t, \quad \text{and} \quad \nabla g_2(\bar{x}) = (1, 2)^t.
\]

Thus, \(\nabla f(\bar{x}) + u_1 \nabla g_1(\bar{x}) + u_2 \nabla g_2(\bar{x}) = 0\) holds true by letting \(u_1 = 1/3\) and \(u_2 = 2/3\), so that the KKT conditions are satisfied at \(\bar{x}\). Noting that \(f, g_1, \text{and } g_2\) are convex, we have that \(\bar{x}\) is indeed an optimal solution by the consequent sufficiency of the KKT conditions.

Now, let us check whether the KKT conditions hold true at the point \(\hat{x} = (0, 0)^t\). Here, \(I = \{3, 4\}\), so that we must have \(u_1 = u_2 = 0\) to satisfy complementary slackness. Note that

\[
\nabla f(\hat{x}) = (-6, -4)^t, \quad \nabla g_3(\hat{x}) = (-1, 0)^t, \quad \text{and} \quad \nabla g_4(\hat{x}) = (0, -1)^t.
\]

Figure B.6 The KKT conditions.
Thus, \( \nabla f(\hat{x}) + u_3 \nabla g_3(\hat{x}) + u_4 \nabla g_4(\hat{x}) = 0 \) holds true only by letting \( u_3 = -6 \) and \( u_4 = -4 \), violating the nonnegativity of the Lagrange multipliers. This shows that \( \hat{x} \) is not a KKT point and hence could not be a candidate for an optimal solution.

In Figure B.6, the gradients of the objective function and the binding constraints are illustrated for both \( \bar{x} \) and \( \hat{x} \). Note that \( -\nabla f(\bar{x}) \) lies in the cone spanned by the gradients of the binding constraints at \( \bar{x} \), whereas \( -\nabla f(\hat{x}) \) does not lie in the corresponding cone. Indeed, the KKT conditions for a problem having inequality constraints could be interpreted geometrically as follows. A vector \( \bar{x} \) is a KKT point if and only if \( -\nabla f(\bar{x}) \) lies in the cone spanned by the gradients of the binding constraints at \( \bar{x} \).

Let Problem P be as defined above, where all objective and constraint functions are continuously twice differentiable, and let \( \bar{x} \) be a KKT solution having associated Lagrange multipliers \( (\bar{u}, \bar{v}) \). Define the (restricted) Lagrangian function \( L(x) = f(x) + \bar{u}'g(x) + \bar{v}'h(x) \), and let \( \nabla^2 L(\bar{x}) \) denote its Hessian at \( \bar{x} \). Let \( C \) denote the cone \( \{ d : \nabla g_i(\bar{x})'d = 0 \text{ for all } i \in I^+, \nabla h_i(\bar{x})'d = 0 \text{ for all } i = 1, \ldots, \ell \} \), where \( I^+ = \{ i \in \{1, \ldots, m\} : \bar{u}_i > 0 \} \) and \( I^0 = \{1, \ldots, m\} - I^+ \). Then we have the following second-order sufficient conditions holding true: If \( \nabla^2 L(\bar{x}) \) is positive definite on \( C \), that is, \( d'\nabla^2 L(\bar{x})d > 0 \) for all \( d \in C, d \neq 0 \), then \( \bar{x} \) is a strict local minimum for Problem P. We also remark that if \( \nabla^2 L(x) \) is positive semidefinite for all feasible \( x \) [respectively, for all feasible \( x \) in \( N_\varepsilon(\bar{x}) \) for some \( \varepsilon > 0 \)], then \( \bar{x} \) is a global (respectively, local) minimum for P.

Conversely, suppose that \( \bar{x} \) is a local minimum for P, and let the gradients \( \nabla g_i(\bar{x}), i \in I, \nabla h_i(\bar{x}), i = 1, \ldots, \ell \) be linearly independent, where \( I = \{ i \in \{1, \ldots, m\} : g_i(\bar{x}) = 0 \} \). Define the cone \( C \) as stated above for the second-order sufficiency conditions. Then \( \bar{x} \) is a KKT point having associated Lagrange multipliers \( (\bar{u}, \bar{v}) \). Moreover, defining the (restricted) Lagrangian function \( L(x) = f(x) + \bar{u}'g(x) + \bar{v}'h(x) \), the second-order necessary condition is that \( \nabla^2 L(\bar{x}) \) is positive semidefinite on \( C \).

### B.4 Lagrangian Duality

Given a nonlinear programming problem, called the primal problem, there exists a problem that is closely associated with it, called the Lagrangian dual problem. These two problems are given below.
Primal Problem P: Minimize $f(x)$
subject to $g_i(x) \leq 0$ for $i = 1, \ldots, m$
$h_i(x) = 0$ for $i = 1, \ldots, \ell$
$x \in X,$

where $f$, $g_i$, and $h_i: \mathbb{R}^n \to \mathbb{R}$ and $X$ is a nonempty set in $\mathbb{R}^n$. Let $g$ and $h$ be the $m$- and $\ell$-vector functions whose $i$th components are, respectively, $g_i$ and $h_i$.

Lagrangian Dual Problem D: Maximize $\theta(u, v)$
subject to $u \geq 0,$

where $\theta(u, v) = \inf\{f(x) + \sum_{i=1}^m u_i g_i(x) + \sum_{i=1}^\ell v_i h_i(x) : x \in X\}$. Here the vectors $u$ and $v$ belong to $\mathbb{R}^m$ and $\mathbb{R}^\ell$, respectively. The $i$th component $u_i$ of $u$ is referred to as the dual variable or Lagrange/Lagrangian multiplier associated with the constraint $g_i(x) \leq 0$, and the $i$th component $v_i$ of $v$ is referred to as the dual variable or Lagrange/Lagrangian multiplier associated with the constraint $h_i(x) = 0$. It may be noted that $\theta$ is a concave function, even in the absence of any convexity or concavity assumptions on $f$, $g_i$, or $h_i$, or convexity of the set $X$.

We summarize below some important relationships between the primal and dual problems:

1. If $x$ is feasible to Problem P and if $(u, v)$ is feasible to Problem D, then $f(x) \geq \theta(u, v)$. Thus,

$$\inf\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\} \geq \sup\{\theta(u, v) : u \geq 0\}.$$

This result is called the weak duality theorem.

2. If $\sup\{\theta(u, v) : u \geq 0\} = \infty$, then there exists no point $x \in X$ such that $g(x) \leq 0$ and $h(x) = 0$, so that the primal problem is infeasible.

3. If $\inf\{f(x) : g(x) \leq 0, h(x) = 0, x \in X\} = -\infty$, then $\theta(u, v) = -\infty$ for each $(u, v)$ with $u \geq 0$.

4. If there exists a feasible $x$ to the primal problem and a feasible $(u, v)$ to the dual problem such that $f(x) = \theta(u, v)$, then $x$ is an optimal solution to Problem P and $(u, v)$ is an optimal solution to Problem D. Furthermore, the complementary slackness condition $u_i g_i(x) = 0$ for $i = 1, \ldots, m$ holds true.

5. Suppose that $X$ is convex, that $f$, $g_i: \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$ are convex, and that $h$ is of the form $h(x) = Ax - b$, where $A$ is an $m \times n$ matrix and $b$ is an $m$-vector. Under a suitable constraint qualification, the optimal objective values of Problems P and D are equal; that is,

$$\inf\{f(x) : x \in X, g(x) \leq 0, h(x) = 0\} = \sup\{\theta(u, v) : u \geq 0\}.$$
Furthermore, if the inf is finite, then the sup is achieved at \((\bar{u}, \bar{v})\) with \(\bar{u} \geq 0\). Also, if the inf is achieved at \(\bar{x}\), then \(u_i g_i(\bar{x}) = 0\) for \(i = 1, \ldots, m\). This result is referred to as the \textit{strong duality theorem}. 


Anderssen, R. S., L. Jennings, and D. Ryan (Eds.), *Optimization*, University of Queensland Press, St. Lucia, Queensland, Australia, 1972.


Bibliography


Bibliography


Cohn, M. Z. (Ed.), *An Introduction to Structural Optimization*, University of Waterloo Press, Waterloo, Ontario, Canada, 1969.


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Han, C. G., P. M. Pardalos, and Y. Ye, “Computational Aspects of an Interior Point Algorithm for Quadratic Programming Problems with Box Constraints,” Computer Science Department, Pennsylvania State University, University Park, PA, 1989.


Bibliography


Bibliography


Karush, W., “Minima of Functions of Several Variables with Inequalities as Side Conditions,” M.S. thesis, Department of Mathematics, University of Chicago, 1939.


Lasdon, L. S., and A. D. Waren, "Generalized Reduced Gradient Software for Linearly and Nonlinearly Constrained Problems," in Design and Implementation of


Powell, M. J. D., "How Bad Are the BFGS and DFP Methods When the Objective Function Is Quadratic?" DAMTP Report 85/NA4, University of Cambridge, Cambridge, 1985c.


Bibliography


Sherali, H. D., and W. P. Adams, “A Hierarchy of Relaxations Between the Continuous
and Convex Hull Representations for Zero-One Programming Problems,” SIAM
Characterizations for Mixed-Integer Zero-One Programming Problems,” Discrete
Sherali, H. D., and W. P. Adams, A Reformulation-Linearization Technique for Solving
Discrete and Continuous Nonconvex Problems, Kluwer Academic, Boston, MA,
1999.
Sherali, H. D., W. P. Adams, and P. Driscoll, “Exploiting Special Structures in
Constructing a Hierarchy of Relaxations for 0–1 Mixed Integer Problems,”
Sherali, H. D., and A. Alameddine, “An Explicit Characterization of the Convex
Envelope of a Bivariate Bilinear Function over Special Polytopes,” in Annals of
Operations Research, Computational Methods in Global Optimization, P. Pardalos
Sherali, H. D., and A. Alameddine, “A New Reformulation-Linearization Algorithm for
the Capacitated Euclidean and \( \ell_p \) Distance Multifacility Location-Allocation
Optimization Methods to Solve Lagrangian Duals of Linear Programs,” Operations
Sherali, H. D., G. Choi, and Z. Ansari, “Limited Memory Space Dilation and Reduction
Algorithms,” Computational Optimization and Applications, 19(1), pp. 55–77,
2001a.
Sherali, H. D., and J. Desai, “On Using RLT to Solve Polynomial, Factorable, and Black-
Box Optimization Problems,” manuscript, Department of Industrial and Systems
Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA,
2004.
Point Mathematical Programming Problem,” Computers and Operations Research,
Sherali, H. D., and B. M. P. Fraticelli, “Enhancing RLT Relaxations via a New Class of
Semidefinite Cuts,” Journal of Global Optimization, special issue in honor of
Professor Reiner Horst, P. M. Pardalos and N. V. Thoai (Eds.), 22(1/4), pp. 233–
261, 2002.
Sherali, H. D. and Ganesan, V. “A Pseudo-global Optimization Approach with
Application to the Design of Containerships,” Journal of Global Optimization,
Sherali, H. D., R. S. Krishnamurthy, and F. A. Al-Khayyal, “An Enhanced Intersection
Cutting Plane Approach for Linear Complementarity Problems,” Journal of
Sherali, H. D., R. S. Krishnamurthy, and F. A. Al-Khayyal, “Enumeration Approach for
Linear Complementarity Problems Based on a Reformulation-Linearization
1998.
Value Method: Application to Solving Lagrangian Relaxations of Linear Programs,”

Sherali, H. D., and C. Lim, “Enhancing Lagrangian Dual Optimization for Linear
Programs by Obviating Nondifferentiability,” *INFORMS Journal on Computing*, to

Sherali, H. D., and D. C. Myers, “The Design of Branch and Bound Algorithms for a
Class of Nonlinear Integer Programs,” *Annals of Operations Research*, Special Issue
on Algorithms and Software for Optimization, C. Monma (Ed.), 5, pp. 463–484,
1986.

Sherali, H. D., and D. C. Myers, “Dual Formulations and Subgradient Optimization
Strategies for Linear Programming Relaxations of Mixed-Integer Programs,”

Sherali, H. D., and S. Sen, “A Disjunctive Cutting Plane Algorithm for the Extreme Point


Sherali, H. D., and C. M. Shetty, “A Finitely Convergent Algorithm for Bilinear
Programming Problems Using Polar Cuts and Disjunctive Face Cuts,” *Mathematical


Sherali, H. D., and C. M. Shetty, *Optimization with Disjunctive Constraints*, Series in
Economics and Mathematical Systems, Publication 181, Springer-Verlag, New
York, NY, 1980c.


Sherali, H. D., and C. M. Shetty, “Nondominated Cuts for Disjunctive Programs and

Analysis for a Perturbation of the Scaling Algorithm for Linear Programming with
Application to the L estimation Problem,” *Naval Research Logistics Quarterly*, 35,

Utility Capacity Planning and Marginal Cost Pricing Problems,” in *Energy Models
and Studies: Studies in Management Science and Systems Series*, B. Lev (Ed.),

Sherali, H. D., A. L. Soyster, and F. H. Murphy, “Stackelberg–Nash–Cournot Equilibria:
1983.

Solar Generation Units in Electric Utility Capacity Expansion Plans,” *Computers

Sherali, H. D., S. Subramanian, and G. V. Loganathan, “Effective Relaxations and
Partitioning Schemes for Solving Water Distribution Network Design Problems to

Programming Problems Using a Reformulation-Linearization Technique,” *Journal of

Solving Nonconvex Quadratic Programming Problems,” *Journal of Global


Bibliography


Bibliography


Ye, Y., "Interior Point Algorithms for Quadratic Programming," Paper Series No. 89–29, Department of Management Sciences, University of Iowa, Iowa City, IA, 1989.


Index

λ-form approximation, 685, 741
ℓ₁ norm, 334
ℓ₁ penalty function, 487
ℓ_p norm metric, 25
(m + 1)-step process, 433
η-invex function, 234
η-pseudoinvex function, 234
η-quasi-invex function, 234
Abadie's Constraint Qualification, 248
absolute value or ℓ₁ penalty function, 487
acceleration step, 367, 368
active constraints, 177
active nodes, 678
active set approach, 651
active set method, 732
active set strategy, 738, 747
addition of matrices, 753
adjacent almost complementary basic feasible solution, 659
affine, 98, 148, 752
affine combination, 40, 752
affine hull, 42
affine hulls, 752
affine independence, 751
affine manifold, 150
affine scaling variant, 634
affinely independent, 43, 751
Aitken double sweep method, 365
ALAG penalty function, 500
algorithm, 317
algorithmic map, 317
almost complementary basic feasible solution, 659
alternative optimal solutions, 2, 124
AMPL, 36
approximating problem, 687
approximating the separable problem, 684
approximation programming, 650
Armijo’s inexact line search, 392
Armijo’s Rule, 362
Arrow–Hurwicz–Uzawa constraint qualification, 254
artificial variables, 83
ascent direction, 283
aspiration level, 21
 augmented Lagrangian, 534
 augmented Lagrangian penalty function (ALAG), 471, 490, 491, 495
 augmented Lagrangian penalty methods, 485
 auxiliary function, 471, 502
 average convergence rates, 341
 average direction strategy, 441, 467
 average rates of convergence, 332
 ball, 760
 BARON, 36
 barrier function methods, 501, 503, 508, 512, 536
 barrier problem, 501
 basic feasible solution, 68
 basic vector, 603
 basis, 77, 752
 BFGS update, 417
 big-M method, 83
 bilinear program, 226, 645
 bilinear programming problem, 657
 bimatrix game, 726
 bounding and scaling, 29
 box-step method, 401
 branch-and-bound algorithm, 678, 738
 branching variable, 679, 680
 Broyden family updates, 416
 Broyden–Fletcher–Goldfarb–Shanno update, 416
 bundle algorithms, 467
 bundle methods, 442
 canonical form, 91
 Carathéodory theorem, 43
 Cauchy point, 402

843
Cauchy sequence, 761
Cauchy's method, 384
central path, 511
chance constrained problems, 35
chance constraint, 21, 22
characterization of optimality, 86
chemical process optimization, 36
children hyperrectangles, 679
choice of step sizes, 439
Cholesky factor, 759
Cholesky factorization, 154, 419, 758
classical optimization techniques, 165
closed, 45, 761
closed half-spaces, 52
closed interval, 760
closed maps, 321
closed maps and convergence, 319
closed sets, 761
closedness of the line search
algorithmic map, 363
closest-point theorem, 50
closure, 45, 761
closure points, 761
cofactor, 754
column vector, 751
compact, 762
compact set, 45, 762
comparison among algorithms, 329
complementarity constraints, 657
complementary basic, 660
complementary basic feasible solution, 657
complementary cone, 657
complementary feasible solution, 657
complementary slack solutions, 86
complementary slackness condition, 86, 183, 191, 772, 773
complementary update, 419
complementary variables, 656
complexity, 122
complexity analysis, 516
component, 751
composite dual, 313
composite map, 324, 326
composition of mappings, 324
computational comparison of reduced
gradient-type methods, 653
computational difficulties associated
with barrier functions, 507
computational difficulties associated
with penalty functions, 481
computational effort, 330
concave, 98, 767
condition number, 389
cone of attainable directions, 242
cone of feasible directions, 174, 242
cone of improving directions, 174
cone of interior directions, 242
cone of tangents, 93, 237, 238, 250
cone spanned, 752
cone spanned by a finite number of
vectors, 765
conjugate, 403
conjugate directions, 402, 407
conjugate dual problem, 313
conjugate functions, 312
conjugate gradient methods, 402, 422, 420
conjugate subgradient algorithm, 534
conjugate weak duality theorem, 313
CONOPT, 36
constraint qualifications, 189, 773, 241
constraint-factor product inequalities, 682
continuity of convex functions, 100
continuous at \( \bar{x} \), 762
continuous functions, 762
continuous optimal control, 6
contour, 3
control problem, 311
control vector, 4
converge, 319
converge to the limit point, 761
convergence, 331
convergence analysis for the RLT
algorithm, 680
convergence analysis of the gradient
projection method, 599
convergence analysis of the method of
Zoutendijk, 557
convergence analysis of the quadratic
programming complementary
pivoting algorithm, 671
convergence analysis of the steepest
descent algorithm, 392
convergence of algorithms, 318, 326
convergence of conjugate direction
methods, 432
convergence of Newton's method, 359
convergence of the bisection search
method, 357
convergence of the complementary
pivoting algorithm, 663
Index

convergence of the cutting plane algorithm, 338
deflected negative gradient, 335
deflecting the gradient, 389
degenerate, 418
degree of difficulty, 717, 720
dense equation systems, 757
derivative-free line search methods, 354
descent direction, 167, 174
descent function, 321, 323
design variables, 17
determinant of a matrix, 754
diadic product, 736
diagonal matrix, 755
diagonalization process, 755
dichotomous search, 347
dichotomous search method, 348
differentiability, 277
differentiable, 109
differentiable at $\mathbf{x}$, 763
differentiable functions, 763
differentiable quasiconvex functions, 137
dimension of the set, 94
direct optimization techniques, 99
direction, 66, 767
direction of descent, 155
direction of steepest ascent, 284
directional derivative, 102
direction-finding routine for a convergent variant of the gradient projection method, 601
direction-finding subproblem, 571
discrepancy index, 679
discrete control problem, 5
discrete optimal control, 4
distance function, 149
distinct directions, 67
dog-leg trajectory, 402
Dorn's dual quadratic program, 300
dual, 84
dual feasibility condition, 191
dual feasibility conditions, 183
dual function: properties, 278
dual geometric program, 716
dual problem, 258, 299
dual update, 419
dual variables, 258
duality gap, 264, 269
duality in linear programming, 84
duality theorems, 263
dualization, 258
direct optimization techniques, 99
dimension of the set, 94
directional derivative, 102
direction-finding routine for a convergent variant of the gradient projection method, 601
differentiable functions, 763
convex, 40, 98, 765, 767
covariance matrix, 20
created response surface technique, 533
criterion function, 2
cumulative distribution function, 149
curvilinear directions, 180
cutting plane, 441, 442
cutting plane algorithm, 338
cutting plane method, 289, 290, 337
cyclic coordinate method, 366, 365
cycling, 80
data fitting, 36
Davidon-Fletcher-Powell (DFP)
method, 408, 414
decrease of the cutting plane algorithm, 338
deflected negative gradient, 335
deflecting the gradient, 389
degenerate, 418
degree of difficulty, 717, 720
dense equation systems, 757
derivative-free line search methods, 354
descent direction, 167, 174
descent function, 321, 323
design variables, 17
determinant of a matrix, 754
diadic product, 736
diagonal matrix, 755
diagonalization process, 755
dichotomous search, 347
dichotomous search method, 348
differentiability, 277
differentiable, 109
differentiable at $\mathbf{x}$, 763
differentiable functions, 763
differentiable quasiconvex functions, 137
dimension of the set, 94
direct optimization techniques, 99
direction, 66, 767
direction of descent, 155
direction of steepest ascent, 284
directional derivative, 102
direction-finding routine for a convergent variant of the gradient projection method, 601
direction-finding subproblem, 571
discrepancy index, 679
discrete control problem, 5
discrete optimal control, 4
distance function, 149
distinct directions, 67
dog-leg trajectory, 402
Dorn's dual quadratic program, 300
dual, 84
dual feasibility condition, 191
dual feasibility conditions, 183
dual function: properties, 278
dual geometric program, 716
dual problem, 258, 299
dual update, 419
dual variables, 258
duality gap, 264, 269
duality in linear programming, 84
duality theorems, 263
dualization, 258
convex combination, 40, 752, 765
covex cone, 41, 62, 63
covex envelope, 151, 736
covex extension, 159
covex functions, 767, 769
covex hull, 40, 42, 752
covex programming problems, 125
covex quadratic programming, 667
covex sets, 765
covexity at a point, 145
covex-simplex method, 613, 705
cordinate vector, 751
copositive, 665
copositive-plus, 665
corrector step, 519
Cottle's constraint qualification, 244, 248
covariance matrix, 20
created response surface technique, 533
criterion function, 2
cumulative distribution function, 149
curvilinear directions, 180
cutting plane, 441, 442
cutting plane algorithm, 338
cutting plane method, 289, 290, 337
cyclic coordinate method, 366, 365
cycling, 80
data fitting, 36
Davidon-Fletcher-Powell (DFP)
method, 408, 414
deflected negative gradient, 335
deflecting the gradient, 389
degenerate, 418
degree of difficulty, 717, 720
dense equation systems, 757
derivative-free line search methods, 354
descent direction, 167, 174
descent function, 321, 323
design variables, 17
determinant of a matrix, 754
diadic product, 736
diagonal matrix, 755
diagonalization process, 755
dichotomous search, 347
dichotomous search method, 348
differentiability, 277
differentiable, 109
differentiable at $\mathbf{x}$, 763
differentiable functions, 763
differentiable quasiconvex functions, 137
dimension of the set, 94
direct optimization techniques, 99
direction, 66, 767
direction of descent, 155
direction of steepest ascent, 284
directional derivative, 102
direction-finding routine for a convergent variant of the gradient projection method, 601
direction-finding subproblem, 571
discrepancy index, 679
discrete control problem, 5
discrete optimal control, 4
distance function, 149
distinct directions, 67
dog-leg trajectory, 402
Dorn's dual quadratic program, 300
dual, 84
dual feasibility condition, 191
dual feasibility conditions, 183
dual function: properties, 278
dual geometric program, 716
dual problem, 258, 299
dual update, 419
dual variables, 258
duality gap, 264, 269
duality in linear programming, 84
duality theorems, 263
dualization, 258
dualized, 286
eccentricity ratio, 12
economic interpretations, 275
effect of near-binding constraints, 556
effective domain, 159
efficient, 31
efficient portfolio, 31
eigenvalue, 755
eigenvector, 755
either-or constraints, 159
elastic constraints, 28
electrical networks, 13
element, 751
empty set, 759
ending node, 14
epigraph, 104
equality constraint, 2
equilibrium problems, 36
error estimations, 694
error function, 391
Euclidean, 25, 334
Euclidean norm, 752
exact absolute value penalty methods, 485
exact penalty function, 485, 487, 491
expanding subspace property, 405
explicitly quasiconvex, 139
exploratory search, 368
exterior penalty function method, 475, 479
extreme direction, 65, 67, 70
extreme directions: existence, 74
extreme point: initial, 82
extreme points, 65, 67
extreme points: existence, 69
factorable programming problems, 748
factorizations, 331
Farkas’s theorem, 767
fathom, 678
feasible direction, 174, 538, 561
feasible direction algorithms, 537
feasible direction method, 129
feasible direction method of Topkis and Veinott, 561
feasible region, 2, 75
feasible solution, 2, 124
Fibonacci search, 351
Fibonacci search method, 353
first-order (Taylor series), 109
first-order condition, 167
first-order linear programming approximation, 193
first-order variants of the reduced gradient method, 620
fixed-metric methods, 421
fixed-time continuous control problem, 6
forest harvesting problems, 36
formulation, 26
fractional programming algorithm, 706
fractional programming problem, 22
Frank-Wolfe method, 631, 733
Frisch’s logarithmic barrier function, 502, 510
Fritz John conditions, 199
Fritz John necessary conditions, 199
Fritz John necessary optimality conditions, 772
Fritz John optimality conditions, 184
Fritz John point, 184
Fritz John sufficient conditions, 187, 203
Frobenius norm, 754
full rank, 754
functionally convex, 139
functions, 762
GAMS, 36
gauge function, 149
Gaussian pivot matrix, 757
Gauss-Seidel iterations, 366
Gauss-Seidel method, 366
generality, 329
generalizations of a convex functions, 134
generalized reduced gradient (GRG) method, 602, 612, 651
geometric convergence, 331
geometric interpretation for the absolute value penalty function, 489
geometric interpretation for the augmented Lagrangian penalty function, 494
geometric interpretation of improving feasible directions, 539
geometric interpretation of penalty functions, 473
global convergence, 399
global minimum, 166
global optimal solution, 124
global optimization approach for
global optimization approaches, 667
globally convergent variant using the
golden section method, 348, 350
Gordan’s theorem, 56, 767
gradient, 40, 763
gradient deflection methods, 421
gradient method, 302, 384
gradient projection method, 589
gradient projection method (linear constraints), 593
gradient vector, 109
Gram–Schmidt procedure, 377
graph, 104
Gordon’s theorem, 56, 767
globally convergent variant using the penalty as a merit function, 581
goal programming, 29
gradient method, 302, 384
gradient projection method, 589
gradient projection method (linear constraints), 593
gradient vector, 109
Gram–Schmidt procedure, 377
graph, 104
greatest lower bound, 760
GRG2, 36
grid points, 346
grid point generation procedure, 698
guidelines for model construction, 26
half-space, 40, 765
H-conjugate, 403
Hessian matrix, 112, 763
higher-order variants of predictor-corrector methods, 535
Hooke and Jeeves method, 368
Hooke and Jeeves method using line searches, 369
Hooke and Jeeves method with discrete steps, 372
Householder transformation matrix, 757
hypercube method, 401
hyperplane, 40, 52, 765
hyperrectangle, 675
hypograph, 104
identity matrix, 753
ill-conditioned Hessian, 482
ill-conditioning, 389, 481
ill-conditioning effect, 507
implicit function theorem, 233, 234
improper separation, 53
improving direction, 174
improving feasible direction, 539, 541
incumbent solution, 678
indefinite, 114, 756
inequality constraint, 2
inequality constraints in the ALAG, 499
inexact line searches, 362
infimum, 760
infinitely differentiable function, 116
inflection point, 172
inner product, 752
inner representation, 72
interior, 45, 760
interior penalty function methods, 506
intersection, 760
interval of uncertainty, 345
inventory, 30
inverse matrix, 754
inverse of a matrix, 757
involutory matrix, 758
isolated local optimal solution, 124
iteration, 317
Jacobian matrix, 184, 191, 763
jamming, 560
journal bearing assembly, 11
journal design problem, 11
Kantorovich inequality, 448
Karmarkar’s algorithm, 634
Karush–Kuhn–Tucker necessary conditions, 206, 218
Karush–Kuhn–Tucker sufficient conditions, 197, 213, 216
Kelley’s cutting plane algorithm, 339
Kirchhoff’s loop law, 15
Kirchhoff’s node law, 15
Kuhn–Tucker’s constraint qualification, 244, 248
Lagrange, 183
Lagrange multipliers, 191, 479, 507, 772
Lagrange–Newton approach, 576, 650
Lagrangian decomposition strategy, 287, 314
Lagrangian dual function, 258
Lagrangian dual problem, 257, 259, 775
Lagrangian dual subproblem, 258
Lagrangian duality, 775
Lagrangian duality theorem, 16
Lagrangian function, 211, 269
Lagrangian multipliers, 16, 183, 191, 258, 772, 773
layering strategy, 287
leading principal submatrix, 771
least distance programming problem, 232
least lower bound, 678
least squares estimation problems, 36
least upper bound, 760
Lemke's complementary pivoting algorithm, 659
level set, 99
level surface, 136
Levenberg–Marquardt, 398
Levenberg–Marquardt method, 400
LGO, 36
limit inferior, 761
limit superior, 761
limited memory, 465
line search, 542
line search problem, 128
line search using derivatives, 357
line search without using derivatives, 344
line segment, 40, 46, 765
linear, 332, 752
linear combination, 40, 752
linear complementary problem, 656
linear convergence rate, 331
linear fractional programming, 703
linear fractional programming problems, 703
linear hull, 42, 752
linear independence, 751
linear independence constraint qualification, 244
linear program, 2, 41, 75, 298, 509
linear subspace, 93, 150
linearization step/phase, 668
linearly independent, 751
LINGO, 36
LINDO, 756
LINSV, 434
Lipschitz condition, 434
Lipschitz continuous, 392
local optimal solution, 124
location of facilities, 24
location problem, 25
location-allocation problem, 25, 36
lower semicontinuity, 101, 762
lower semicontinuous at \( \bar{x} \), 763
lower-level set, 99, 136
LU factorization, 668, 756
LU-decomposition, 154
Mangasarian–Fromovitz constraint qualification, 251
Maratos effect, 587, 650
marginal rates of change, 275
master program, 291
mathematical economics, 36
mathematical tractability, 26
MATLAB, 756
matrices, 751, 753
matrix factorizations, 756
matrix multiplication, 753
maxima semicontinuous functions, 762
maximizing a convex function, 133
mean, 20
mean value theorem, 764
mechanical design, 11
memoryless quasi-Newton method, 429
merit function, 569, 581
merit function SQP algorithm, 583
method of multipliers, 495, 538
minima and maxima of convex functions, 123
minima of semicontinuous functions, 762
minimizing a concave objective function, 658
minimizing a convex function, 124
minimizing along independent directions, 327
MINOS, 36, 622
mixed penalty-barrier auxiliary function, 527, 533
mixed-integer zero-one bilinear programming problem, 657
modified barrier method, 536
multidimensional search using derivatives, 384
multidimensional search without using derivatives, 365
 multinominal terms, 737
multiple pricing, 622
multiplication by a scalar, 751
multiplier penalty function, 490, 491
multiset, 737
\( n \)-dimensional Euclidean space, 751
near-binding constraints, 556, 650
near-feasible, 295
near-optimal solution, 296
necessary optimality conditions, 166
negative definite, 756
negative entropy function, 531
negative semidefinite, 113, 756
neighborhood, 760, 765
NEOS server for optimization, 466
Newton’s method, 358, 395
Newton-Raphson method, 395, 577
Newton-Raphson scheme, 613
node potential, 15
node-branch incidence matrix, 14
nonbasic variables, 621
nonbasic vector, 603
nonconvex cone, 63
nonconvex problems, 474
nonconvex quadratic programming, 667
nonconvex quadratic programming problem, 675
nondegeneracy assumption, 603
nonlinear complementary problem, 727
nonlinear equality constraints, 553
nonlinear equation systems, 420
nonlinear equations, 395
nonnegative orthant, 157
nonsingular, 754
norm of a matrix, 754
normal, 40, 52
normalization constraint, 541, 715
n-step asymptotic results, 434
objective function, 2, 75
ones, 657
open, 45, 760
open half-space, 52, 765
open half-spaces,
open interval, 760
open sets, 760
operational variables, 17
optimal control problems, 4, 34
optimal solution, 2, 124
optimality conditions, 772 (also, see Fritz John and Karush-Kuhn-Tucker)
optimality conditions in linear programming, 75
order of complexity, 517
order of convergence, 331
order-two convergence of the method of Newton, 395
origin node, 14
orthogonal, 752
orthogonal complement, 93
orthogonal matrix, 753, 755
orthogonality constraint, 715
orthogonalization procedure, 377
orthonormal matrix, 755
outer approximation, 292
outer linearization, 292
outer product, 736
outer representation, 72
outer-linearization method, 289, 290
parallel tangents, 458
parallelogram law, 50
parameter-free barrier function method, 528
parameter-free methods, 533
parameter-free penalty function method, 528
PARTAN, 428
partial conjugate gradient method, 433
partial quasi-Newton method, 408, 434
partition, 679
partitioned matrices, 753
path-following procedure, 513
pattern search, 368, 428
pattern search step, 368
penalty function, 470, 471
penalty function methods, 484
penalty problem, 476
penalty successive linear programming algorithm, 569
percentage test line search map, 445
perfect line searches, 418
permutation matrix, 753, 756
perturbation function, 150, 260, 273, 474
perturbed KKT system, 511
perturbed primal problems, 293
Phase I, 83
pivoting, 81
PLU decomposition, 757
PLU factorization, 756
P-matrices, 744
polar cone, 63
polar set, 88
polarity, 62
Polyak-Kelly cutting plane method, 442, 467
polyhedral cone, 765
polyhedral set, 40, 65, 765
polynomial function, 737
polynomial in the size of the problem, 517
polynomial programming problem, 655, 668, 737
polynomial-time complexity, 517
polynomial-time interior point algorithms, 509
polytope, 43
portfolio selection problem, 31
positive definite, 756
positive definite secant update, 417
positive semidefinite, 113, 115, 756
positive subdefinite, 157
posynomial, 712
posynomial programming problems, 712
Powell–Symmetric Broyden (PSB) update, 453
practical line search methods, 360
precision, 329
preconditioning matrix, 428
predictor step, 519
predictor-corrector algorithms, 519
predictor-corrector approach, 519
predictor-corrector variants, 535
preparational effort, 330
pricing phase, 622
primal feasibility condition, 183, 191
primal feasible solutions, 296
primal methods, 537
primal problem, 84, 257, 258, 775
primal solution, 293
primal-dual path-following, 537
principal pivoting method, 659, 725, 746
principal submatrix, 771
problems having inequality and equality constraints, 197, 245
problems having inequality constraints, 174
problems having nonlinear inequality constraints, 547
production-inventory, 5
programming with recourse, 35
projected, 576
projected Hessian matrix, 624
projected Lagrangian, 650
projected Lagrangian approach, 576
projecting, 435
projecting the gradient in the presence of nonlinear constraints, 599
projection matrix, 589
proper convex function, 159
proper supporting hyperplane, 57
properly separate, 53
pseudoconcave, 142
pseudoconvex functions, 142, 768, 770
pseudoconvexity at \( \bar{x} \), 146
pure Broyden update, 417
purified, 517
q-order convergence rates, 341
QR decomposition, 759
QR factorization, 757
QRP factorization, 757
quadratic, 332
quadratic approximation, 153
quadratic assignment program, 223
quadratic case, 412, 423
quadratic form, 755
quadratic functions, 405
quadratic penalty function, 480
quadratic programming, 15, 21, 24, 298, 299, 311, 576
quadratic rate of convergence, 331
quadratic-fit algorithm, 446
quadratic-fit line search, 361
quasiconcave, 135
quasiconvex, 768
quasiconvex functions, 134, 771
quasiconvexity at \( \bar{x} \), 145
quasimonotone, 135
quasi-Newton, 402
quasi-Newton approximations, 581
quasi-Newton condition, 416
quasi-Newton procedures, 407
r-(root)-order, 332
rank of a matrix, 754
rank-one correction, 467
rank-one correction algorithm, 455
rank-two correction procedure, 408
rank-two correction procedures, 464
rank-two DFP update, 416
ray termination, 660, 662
real Euclidean space, 4
recession direction, 66
rectilinear, 25
recursive linear programming, 568
recursive programming approaches, 576
reduced cost coefficients, 81
reduced gradient, 605
reduced gradient methods, 653
reduced gradient algorithm, 602, 605
reformulation
linearization/convexification
Index

technique (RLT), 667, 668, 676, 712, 736, 748
reformulation step/phase, 668, 676
regular, 205
relative interior, 94, 760
relative minima, 124
reliability, 329, 330
remainder term, 109
representability, 26
representation of polyhedral sets, 72
representation theorem, 72
response surface methodology, 36
restarting conjugate gradient methods, 430
restricted basis entry rule, 687
restricted Lagrangian function, 212
restricted step methods, 400
risk aversion constant, 23
risk aversion model, 22
river basin, 17
RLT, 667, 668, 675, 712, 736, 746
RLT Algorithm to Solve Problem NQP, 679
RLT constraints, 735
RLT constraints: cubic, 683
RLT variables: cubic, 734
robustness, 330
rocket launching, 7
r-order convergence rates, 341
Rosenbrock’s method, 376
Rosenbrock’s method using line searches, 382
Rosenbrock’s method with discrete steps, 382
rounded, 517
row vector, 751
rudimentary SQP algorithm, 579
saddle point, 172, 269
saddle point criteria, 269, 271
saddle point optimality, 269
saddle point optimality conditions, 213, 263
saddle point optimality interpretation, 273
safeguard technique, 360
satisficing criteria, 21
satisficing level, 21
scalar multiplication of a matrix, 753
scale invariant, 330
scale-invariant algorithms, 29
scaling of Quasi-Newton algorithms, 420
Schwartz inequality, 47, 752
secant equation, 416
second-order (Taylor Series), 112, 113
second-order conditions, 167
second-order cone of attainable directions constraint qualification, 252
second-order constraint qualification, 249
second-order functional approximations, 622
second-order necessary and sufficient conditions for constrained problems, 211
second-order necessary condition, 775
second-order rate of convergence, 331
second-order sufficient conditions, 775
second-order variants of the reduced gradient method, 620
self-scaling methods, 420
semidefinite cuts, 683, 737
semidefinite matrices, 756
semidefinite programming, 683, 736, 748
semi-infinite nonlinear programming problems, 235
semi-strictly quasiconvex, 139, 141
sensitivity analyses, 26, 256, 533
sensitivity to parameters and data, 330
separable nonlinear program, 684
separable programming, 684
separate, 53
separating hyperplane, 766
separation of two convex sets, 59
separation of two sets, 52
sequence, 759, 761
sequential linear programming, 568
sequential programming approaches, 576
sequential search procedures, 347
sequential unconstrained minimization technique (SUMT), 484
sets, 759
Sherman–Morrison–Woodbury formula, 419
shifted/modified barrier method, 534
signomial programming problem, 712
signomials, 712
simplex, 43
simplex method, 75, 76, 456, 652
simplex tableau, 80
simultaneous search, 346
Index

single-step procedure, 434
singular-value decomposition, 755
skew symmetric, 746, 753
slack variable, 77, 81
Slater's constraint qualification, 243, 247
soft constraints, 28
software description, 653
solid set, 45, 771
solution, 124
solution procedure, 317
solution set, 318
space dilation, 441, 466
spacel step, 326, 432
spanning vectors, 752
spectral decomposition, 752
square root (of matrix), 755
standard format, 77
state vector, 4
statistical parameter estimation, 36
steepest ascent directions, 283
steepest descent algorithm, 387
steepest descent algorithm with affine scaling, 394
steepest descent method, 384
step bounds, 568
step length, 441
stochastic resource allocation, 20
straight-line directions, 180
strict complementary slackness condition, 499
strict containment, 760
strict convexity at \( \bar{x} \), 145
strict local optimal solution, 124
strict Mangasarian–Fromovitz constraint qualification, 255
strict pseudoconvexity at \( \bar{x} \), 146
strict quasiconvexity at \( \bar{x} \), 145
strict unimodality, 345
strictly concave, 98, 767
strictly convex, 98, 767, 769
strictly positive subdefinite, 157
strictly pseudoconcave, 142
strictly pseudoconvex, 142, 768
strictly quasiconcave, 139
strictly quasiconvex, 139, 141, 768
strictly separate, 53
strictly unimodal, 445
strong duality result, 85
strong duality theorem, 267, 777
strong local optimal solution, 124
strong quasiconvexity, 140
strong quasiconvexity at \( \bar{x} \), 146
strong separation, 54, 61
strongly active constraints, 213
strongly monotone, 727
strongly quasiconcave, 141
strongly quasiconvex, 140, 768
strongly unimodal, 156, 445
structural design, 9
subadditive, 149
subdifferential, 105
subgradient, 105, 103, 279
subgradient deflection, 441
subgradient optimization, 435
subhyperrectangles, 679
suboptimization strategy, 622
subproblem, 290, 291
subsequence, 761
subset, 759
successive linear programming approach, 568, 650
successive quadratic programming approach, 576, 576, 650
sufficient optimality conditions, 168
sum map, 334
sum vector, 751
sup norm, 334
superbasic variables, 621
superlinear, 332
superlinear convergence, 331
superlinearly convergent polynomial time primal-dual path-following methods, 534
support function, 149
support of sets at boundary points, 57
supporting hyperplane, 57, 58, 766
supporting hyperplane method, 340
supremum, 760
surrogate dual problem, 313
surrogate relaxation, 232
symmetric, 753
tableau format of the simplex method, 80
tangential approximation, 292
target value, 443
Taylor series
  first–order, 109
  second–order, 112, 113
Taylor's theorem, 764
terminating the algorithm, 323
three-point pattern, 361, 446, 447
three-term recurrence relationships, 428
Index

tight constraints, 177
Topkis-Veinott's modification of the feasible direction algorithm, 561
trajectory, 4
trajectory constraint function, 5
transportation problem, 25
transpose, 753
transposition, 753
triangularization, 757
truncated Newton methods, 465
trust region methods, 398, 400
trust region parameter, 400
trust region restrictions, 568
trust region subproblem, 400
twice differentiable, 112
twice differentiable at $\bar{x}$, 763
two-bar truss, 9
unbounded, 77
unbounded-infeasible relationship, 84
unconstrained problems, 166
uniform search, 346
unimodal, 156
union, 759
unit lower triangular matrix, 757
unit vector, 751
univariate case, 169
upper semicontinuity, 101, 762
upper-level set, 99, 136
utility functions, 23
value methods, 441
variable dimension, 461
variable metric method, 408
variable target, 441
variable target value methods, 443, 467
variable-metric methods, 421
vector multiplication by a scalar, 751
vector addition, 751
vector function, 762
vectors, 751
vector-valued function, 763
Veinott's supporting hyperplane algorithm, 340
volume algorithm, 467
warehouse location problem, 311
water resources management, 17
weak duality result, 84
weak duality theorem, 263, 776
weakly active constraints, 213
Wolfe's counterexample, 560
Zangwill's constraint qualification, 244
Zangwill's convergence theorem, 321
zero degrees of difficulty, 719
zero matrix, 753
zero vector, 751
zigzagging, 389, 553
zigzagging of the convex-simplex method, 623
zigzagging of the steepest descent method, 387
Z-matrix, 746
Zoutendijk's method, 537
Zoutendijk's method (case of linear constraints), 542
Zoutendijk's method (case of nonlinear inequality constraints), 550