OPTIMIZATION BY VECTOR SPACE METHODS

OPTIMIZATION BY VECTOR SPACE METHODS

SERIES IN DECISION AND CONTROL

Ronald A. Howard

OPTIMIZATION BY VECTOR SPACE METHODS

by David G. Luenberger

INTRODUCTION TO DYNAMIC PROGRAMMING

by George L. Nemhauser

DYNAMIC PROBABILISTIC SYSTEMS (In Preparation)

by Ronald A. Howard

OPTIMIZATION BY VECTOR SPACE METHODS

David G. Luenberger

Stanford University,
Stanford, California

John Wiley & Sons, Inc.
New York — London — Sydney — Toronto

Library of Congress
Catalog Card Number:
68-8716 SBN 471 55359x
Printed in the United
States of America
3 4 5 6 7 8 9 10

To Nancy

PREFACE

This book has evolved from a course on optimization that I have taught at Stanford University for the past five years. It is intended to be essentially self-contained and should be suitable for classroom work or self-study. As a text it is aimed at first- or second-year graduate students in engineering, mathematics, operations research, or other disciplines dealing with optimization theory.

The primary objective of the book is to demonstrate that a rather large segment of the field of optimization can be effectively unified by a few geometric principles of linear vector space theory. By use of these principles, important and complex infinite-dimensional problems, such as those generated by consideration of time functions, are interpreted and solved by methods springing from our geometric insight. Concepts such as distance, orthogonality, and convexity play fundamental roles in this development. Viewed in these terms, seemingly diverse problems and techniques often are found to be intimately related.

The essential mathematical prerequisite is a familiarity with linear algebra, preferably from the geometric viewpoint. Some familiarity with elementary analysis including the basic notions of sets, convergence, and continuity is assumed, but deficiencies in this area can be corrected as one progresses through the book. More advanced concepts of analysis such as Lebesgue measure and integration theory, although referred to in a few isolated sections, are not required background for this book.

Imposing simple intuitive interpretations on complex infinite-dimensional problems requires a fair degree of mathematical sophistication. The backbone of the approach taken in this book is functional analysis, the study of linear vector spaces. In an attempt to keep the mathematical prerequisites to a minimum while not sacrificing completeness of the development, the early chapters of the book essentially constitute an introduction to functional analysis, with applications to optimization, for those having the relatively modest background described above. The mathematician or more advanced student may wish simply to scan Chapters 2, 3, 5, and 6 for review or for sections treating applications and then concentrate on the other chapters which deal explicitly with optimization theory.

The sequencing of the various sections is not necessarily inviolable. Even at the chapter level the reader may wish to alter his order of progress through the book. The course from which this text developed is two quarters (six months) in duration, but there is more material in the text than can be comfortably covered in that period. By reading only the first few sections of Chapter 3, it is possible to go directly from Chapters 1 and 2 to Chapters 5, 7, 8, and 10 for a fairly comprehensive treatment of optimization which can be covered in about one semester. Alternatively, the material at the end of Chapter 6 can be combined with Chapters 3 and 4 for a unified introduction to Hilbert space problems. To help the reader make intelligent decisions regarding his order of progress through the book, sections of a specialized or digressive nature are indicated by an *.

The problems at the end of each chapter are of two basic varieties. The first consists of miscellaneous mathematical problems and proofs which extend and supplement the theoretical material in the text; the second consists of optimization problems which illustrate further areas of application and which hopefully will help the student formulate and solve practical problems. The problems represent a major component of the book, and the serious student will not pass over them lightly.

I have received help and encouragement from many people during the years of preparation of this book. Of great benefit were comments and suggestions of Pravin Varaiya, E. Bruce Lee, and particularly Samuel Karlin who read the entire manuscript and suggested several valuable improvements. I wish to acknowledge the Departments of Engineering-Economic Systems and Electrical Engineering at Stanford University for supplying much of the financial assistance. This effort was also partially supported by the Office of Naval Research and the National Science Foundation. Of particular benefit, of course, have been the faces of puzzled confusion or of elated understanding, the critical comments and the sincere suggestions of the many students who have worked through this material as the book evolved.

DAVID G. LUENBERGER

Palo Alto, California
August 1968

CONTENTS

NOTATION

Sets

If x is a member of the set S, we write $x \in S$. The notation $y \notin S$ means y is not a member of S.

A set may be specified by listing its elements between braces such as $S = \{1, 2, 3\}$ for the set consisting of the first three positive integers. Alternatively, a set S may be specified as consisting of all elements of the set X which have the property P. This is written $S = \{x \in X: P(x)\}$ or, if X is understood, $S = \{x: P(x)\}$.

The *union* of two sets S and T is denoted $S \cup T$ and consists of those elements that are in either S or T.

The *intersection* of two sets S and T is denoted $S \cap T$ and consists of those elements that are in both S and T. Two sets are *disjoint* if their intersection is empty.

If S is defined as a subset of elements of X, the *complement* of S, denoted $\tilde{S}$, consists of those elements of X that are not in S.

A set S is a *subset* of the set T if every element of S is also an element of T. In this case we write $S \subset T$ or $T \supset S$. If $S \subset T$ and S is not equal to T then S is said to be a *proper subset* of T.

Sets of Real Numbers

If a and b are real numbers, $[a, b]$ denotes the set of real numbers x satisfying $a \leq x \leq b$. A rounded instead of square bracket denotes strict inequality in the definition. Thus $(a, b]$ denotes all x with $a < x \leq b$.

If S is a set of real numbers bounded above, then there is a smallest real number y such that $x \leq y$ for all $x \in S$. The number y is called the *least upper bound* or *supremum* of S and is denoted $\sup_{x \in S}(x)$ or $\sup\{x: x \in S\}$ If S is not bounded above we write $\sup_{x \in S}(x) = \infty$. Similarly, the *greatest lower bound* or *infimum* of a set S is denoted $\inf_{x \in S}(x)$ or $\inf\{x: x \in S\}$.

Sequences

A sequence $x_1, x_2, \ldots, x_n, \ldots$ is denoted by $\{x_i\}_{i=1}^{\infty}$ or $\{x_i\}$ if the range of the indices is clear.

Let $\{x_i\}$ be an infinite sequence of real numbers and suppose that there is a real number S satisfying: (1) for every $\varepsilon > 0$ there is an N such that for all $n > N$, $x_n < S + \varepsilon$, and (2) for every $\varepsilon > 0$ and $M > 0$ there is an $n > M$ such that $x_n > S - \varepsilon$. Then S is called the *limit superior* of $\{x_n\}$ and we write $S = \limsup_{n\to\infty} x_n$. If $\{x_n\}$ is not bounded above we write $\limsup x_n = +\infty$. The *limit inferior* of x_n is $\liminf x_n = -\limsup(-x_n)$. If $\limsup x_n = \liminf x_n = S$, we write $\lim x_n = S$.

Functions

The function sgn (pronounced sig-num) of a real variable is defined by

$$\operatorname{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

The Kronecker delta function δ_{ij} is defined by

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i = j \end{cases}$$

The Dirac delta function δ is used occasionally in heuristic discussions. It is defined by the relation

$$\int_a^b f(t)\delta(t)\,dt = f(0)$$

for every continuous function f provided that $0 \in (a, b)$.

If g is a real-valued function of a real variable we write $S = \limsup_{x\to x_0} g(x)$ if: (1) for every $\varepsilon > 0$ there is $\delta > 0$ such that for all x satisfying $|x - x_0| < \delta$, $g(x) < S + \varepsilon$, and (2) for every $\varepsilon > 0$ and $\delta > 0$ there is an x such that $|x - x_0| < \delta$ and $g(x) > S - \varepsilon$. (See the corresponding definitions for sequences.)

If g is a real-valued function of a real variable, the notation $g(x) = O(x)$ means that

$$K = \limsup_{x\to 0} \left|\frac{g(x)}{x}\right|$$

is finite. The notation $g(x) = o(x)$ means that K, above, is zero.

Matrices and Vectors

A vector x with n components is written $x = (x_1, x_2, \ldots, x_n)$, but when used in matrix calculations it is represented as a *column vector*, i.e.,

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

The corresponding *row vector* is

$$x' = \underbrace{x_1 \; x_2 \cdots x_n}$$

An $n \times m$ matrix A with entry a_{ij} in its i-th row and j-th column is written $A = [a_{ij}]$. If $x = (x_1, x_2, \ldots, x_n)$, the product Ax is the vector y with components $y_i = \sum_{j=1}^{n} a_{ij}x_j$, $i = 1, 2, \ldots, m$.

Let $f(x_1, x_2, \ldots, x_n)$ be a function of the n real variables x_i. Then we write f_x for the row vector

$$\underbrace{\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}}.$$

If $F = (f_1, f_2, \ldots, f_m)$ is a vector function of $x = (x_1, \ldots, x_n)$, we write F_x for the $m \times n$ Jacobian matrix $[\partial f_i / \partial x_j]$.

1

INTRODUCTION

1.1 Motivation

During the past twenty years mathematics and engineering have been increasingly directed towards problems of decision making in physical or organizational systems. This trend has been inspired primarily by the significant economic benefits which often result from a proper decision concerning the distribution of expensive resources, and by the repeated demonstration that such problems can be realistically formulated and mathematically analyzed to obtain good decisions.

The arrival of high-speed digital computers has also played a major role in the development of the science of decision making. Computers have inspired the development of larger systems and the coupling of previously separate systems, thereby resulting in decision and control problems of correspondingly increased complexity. At the same time, however, computers have revolutionized applied mathematics and solved many of the complex problems they generated.

It is perhaps natural that the concept of best or optimal decisions should emerge as the fundamental approach for formulating decision problems. In this approach a single real quantity, summarizing the performance or value of a decision, is isolated and optimized (i.e., either maximized or minimized depending on the situation) by proper selection among available alternatives. The resulting optimal decision is taken as the solution to the decision problem. This approach to decision problems has the virtues of simplicity, preciseness, elegance, and, in many cases, mathematical tractability. It also has obvious limitations due to the necessity of selecting a single objective by which to measure results. But optimization has proved its utility as a mode of analysis and is firmly entrenched in the field of decision making.

Much of the classical theory of optimization, motivated primarily by problems of physics, is associated with great mathematicians: Gauss, Lagrange, Euler, the Bernoulis, etc. During the recent development of optimization in decision problems, the classical techniques have been re-examined, extended, sometimes rediscovered, and applied to problems

having quite different origins than those responsible for their earlier development. New insights have been obtained and new techniques have been discovered. The computer has rendered many techniques obsolete while making other previously impractical methods feasible and efficient. These recent developments in optimization have been made by mathematicians, system engineers, economists, operations researchers, statisticians, numerical analysts, and others in a host of different fields.

The study of optimization as an independent topic must, of course, be regarded as a branch of applied mathematics. As such it must look to various areas of pure mathematics for its unification, clarification, and general foundation. One such area of particular relevance is functional analysis.

Functional analysis is the study of vector spaces resulting from a merging of geometry, linear algebra, and analysis. It serves as a basis for aspects of several important branches of applied mathematics including Fourier series, integral and differential equations, numerical analysis, and any field where linearity plays a key role. Its appeal as a unifying discipline stems primarily from its geometric character. Most of the principal results in functional analysis are expressed as abstractions of intuitive geometric properties of ordinary three-dimensional space.

Some readers may look with great expectation toward functional analysis, hoping to discover new powerful techniques that will enable them to solve important problems beyond the reach of simpler mathematical analysis. Such hopes are rarely realized in practice. The primary utility of functional analysis for the purposes of this book is its role as a unifying discipline, gathering a number of apparently diverse, specialized mathematical tricks into one or a few general geometric principles.

1.2 Applications

The main purpose of this section is to illustrate the variety of problems that can be formulated as optimization problems in vector space by introducing some specific examples that are treated in later chapters. As a vehicle for this purpose, we classify optimization problems according to the role of the decision maker. We list the classification, briefly describe its meaning, and illustrate it with one problem that can be formulated in vector space and treated by the methods described later in the book. The classification is not intended to be necessarily complete nor, for that matter, particularly significant. It is merely representative of the classifications often employed when discussing optimization.

Although the formal definition of a vector space is not given until Chapter 2, we point out, in the examples that follow, how each problem

can be regarded as formulated in some appropriate vector space. However, the details of the formulation must, in many cases, be deferred until later chapters.

1. Allocation. In allocation problems there is typically a collection of resources to be distributed in some optimal fashion. Almost any optimization problem can be placed in this broad category, but usually the term is reserved for problems in which the resources are distributed over space or among various activities.

A typical problem of this type is that faced by a manufacturer with an inventory of raw materials. He has certain processing equipment capable of producing n different kinds of goods from the raw materials. His problem is to allocate the raw materials among the possible products so as to maximize his profit.

In an idealized version of the problem, we assume that the production and profit model is linear. Assume that the selling price per unit of product j is p_j, $j = 1, 2, \ldots, n$. If x_j denotes the amount of product j that is to be produced, b_i the amount of raw material i on hand, and a_{ij} the amount of material i in one unit of product j, the manufacturer seeks to maximize his profit

$$p_1 x_1 + p_2 x_2 + \cdots + p_n x_n$$

subject to the production constraints on the amount of raw materials

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &\leq b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &\leq b_2 \\ \vdots \qquad\qquad\qquad\qquad &\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &\leq b_m \end{aligned}$$

and

$$x_1 \geq 0, x_2 \geq 0, \ldots, x_n \geq 0.$$

This type of problem, characterized by a linear objective function subject to linear inequality constraints, is a linear programming problem and is used to illustrate aspects of the general theory of optimization in later chapters.

We note that the problem can be regarded as formulated in ordinary n-dimensional vector space. The vector x with components x_i is the unknown. The constraints define a region in the vector space in which the selected vector must lie. The optimal vector is the one in that region maximizing the objective.

The manufacturing problem can be generalized to allow for nonlinear

objectives and more general constraints. Linearity is destroyed, which may make the solution more difficult to obtain, but the problem can still be regarded as one in ordinary Euclidean n-dimensional space.

*2. **Planning.*** Planning is the problem of determining an optimal procedure for attaining a set of objectives. In common usage, planning refers especially to those problems involving outlays of capital over a period of time such as (1) planning a future investment in electric power generation equipment for a given geographic region or (2) determining the best hiring policy in order to complete a complex project at minimum expense.

As an example, consider a problem of production planning. A firm producing a certain product wishes to plan its production schedule over a period of time in an optimal fashion. It is assumed that a fixed demand function over the time interval is known and that this demand must be met. Excess inventory must be stored at a storage cost proportional to the amount stored. There is a production cost associated with a given rate of production. Thus, denoting $x(t)$ as the stock held at time t, $r(t)$ as the rate of production at time t, and $d(t)$ as the demand at time t, the production system can be described by the equations[1]

$$\dot{x}(t) = r(t) - d(t), \qquad x(0) \text{ given}$$

and one seeks the function r satisfying the inequality constraints

$$\left.\begin{aligned} r(t) &\geq 0, \\ x(0) + \int_0^t [r(\tau) - d(\tau)]\, d\tau = x(t) &\geq 0 \end{aligned}\right\} \qquad \text{for } 0 \leq t \leq T$$

and minimizing the cost

$$J = \int_0^T \{c[r(t)] + h \cdot x(t)\}\, dt$$

where $c[r]$ is the production cost rate for the production level r and $h \cdot x$ is the inventory cost rate for inventory level x.

This problem can be regarded as defined on a vector space consisting of continuous functions on the interval $[0, T]$ of the real line. The optimal production schedule r is then an element of the space. Again the constraints define a region in the space in which the solution r must lie while minimizing the cost.

*3. **Control (or Guidance).*** Problems of control are associated with dynamic systems evolving in time. Control is quite similar to planning;

[1] $\dot{x}(t) \equiv dx(t)/dt$.

indeed, as we shall see, it is often the source of a problem rather than its mathematical structure which determines its category.

Control or guidance usually refers to directed influence on a dynamic system to achieve desired performance. The system itself may be physical in nature, such as a rocket heading for Mars or a chemical plant processing acid, or it may be operational such as a warehouse receiving and filling orders.

Often we seek feedback or so-called closed-loop control in which decisions of current control action are made continuously in time based on recent observations of system behavior. Thus, one may imagine himself as a controller sitting at the control panel watching meters and turning knobs or in a warehouse ordering new stock based on inventory and predicted demand. This is in contrast to the approach described for planning in which the whole series of control actions is predetermined. Generally, however, the terms planning or control may refer to either possibility.

As an example of a control problem, we consider the launch of a rocket to a fixed altitude h in given time T while expending a minimum of fuel. For simplicity, we assume unit mass, a constant gravitational force, and the absence of aerodynamic forces. The motion of a rocket being propelled vertically is governed by the equations

$$\ddot{y}(t) = u(t) - g$$

where y is the vertical height, u is the accelerating force, and g is the gravitational force. The optimal control function u is the one which forces $y(T) = h$ while minimizing the fuel expenditure $\int_0^T |u(t)|\, dt$.

This problem too might be formulated in a vector space consisting of functions u defined on the interval $[0, T]$. The solution to this problem, however, is that $u(t)$ consists of an impulse at $t = 0$ and, therefore, correct problem formulation and selection of an appropriate vector space are themselves interesting aspects of this example. Problems of this type, including this specific example, are discussed in Chapter 5.

4. Approximation. Approximation problems are motivated by the desire to approximate a general mathematical entity (such as a function) by one of simpler, specified form. A large class of such approximation problems is important in numerical analysis. For example, suppose we wish, because of storage limitations or for purposes of simplifying an analysis, to approximate a function, say $x(t)$, over an interval $[a, b]$ of the real line by a polynomial $p(t)$ of order n. The best approximating polynomial p minimizes the error $e = x - p$ in the sense of some criterion. The choice of criterion determines the approximation. Often used criteria are:

1. $\int_a^b e^2(t)\,dt$
2. $\max_{a \le t \le b} |e(t)|$
3. $\int_a^b |e(t)|\,dt.$

The problem is quite naturally viewed as formulated in a vector space of functions over the interval $[a, b]$. The problem is then viewed as finding a vector from a given class (polynomials) which is closest to a given vector.

5. Estimation. Estimation problems are really a special class of approximation problems. We seek to estimate some quantity from imperfect observations of it or from observations which are statistically correlated but not deterministically related to it. Loosely speaking, the problem amounts to approximating the unobservable quantity by a combination of the observable ones. For example, the position of a random maneuvering airplane at some future time might reasonably be estimated by a linear combination of past measurements of its position.

Another example of estimation arises in connection with triangulation problems such as in location of forest fires, ships at sea, or remote stars. Suppose there are three lookout stations, each of which measures the angle of the line-of-sight from the station to the observed object. The situation is illustrated in Figure 1.1. Given these three angles, what is the best estimate of the object's location?

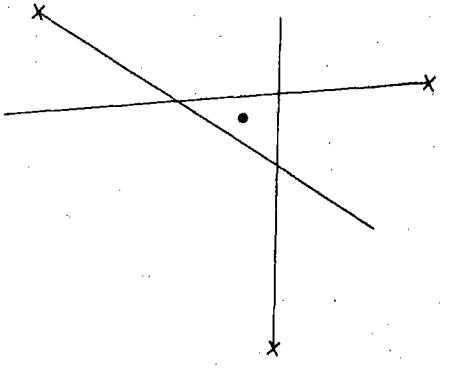

Figure 1.1 A triangulation problem

To formulate the problem completely, a criterion must be precisely prescribed and hypotheses specified regarding the nature of probable

measurement errors and probable location of the object. Approaches can be taken that result in a problem formulated in vector space; such problems are discussed in Chapter 4.

6. Games. Many problems involving a competitive element can be regarded as games. In the usual formulation, involving two players or protagonists, there is an objective function whose value depends jointly on the action employed by both players. One player attempts to maximize this objective while the other attempts to minimize it.

Often two problems from the categories discussed above can be competitively intermixed to produce a game. Combinations of categories that lead to interesting games include: allocation-allocation, allocation-control, control-control, and estimation-control.

As an example, consider a control-control game. Most problems of this type are of the pursuer-evader type such as a fighter plane chasing a bomber. Each player has a system he controls but one is trying to maximize the objective (time to intercept for instance) while the other is trying to minimize the objective.

As a simpler example, we consider a problem of advertising or campaigning which is essentially an allocation-allocation game.[2] Two opposing candidates, A and B, are running for office and must plan how to allocate their advertising resources (A and B dollars, respectively) among n distinct geographical areas. Let x_i and y_i denote, respectively, the resources committed to area i by candidates A and B. We assume that there are currently a total of u undecided votes of which there are u_i undecided votes in area i. The number of votes going to candidates A and B from area i are assumed to be

$$\frac{x_i u_i}{x_i + y_i}, \qquad \frac{y_i u_i}{x_i + y_i}$$

respectively. The total difference between the number of votes received by A and by B is then

$$\sum_{i=1}^{n} \frac{x_i - y_i}{x_i + y_i} u_i.$$

Candidate A seeks to maximize this quantity while B seeks to minimize it.

This problem is obviously finite dimensional and can be solved by ordinary calculus in a few lines. It is illustrative, however, of an interesting class of game problems.

[2] This problem is due to L. Friedman [57].

1.3 The Main Principles

The theory of optimization presented in this book is derived from a few simple, intuitive, geometric relations. The extension of these relations to infinite-dimensional spaces is the motivation for the mathematics of functional analysis which, in a sense, often enables us to extend our three-dimensional geometric insights to complex infinite-dimensional problems. This is the conceptual utility of functional analysis. On the other hand, these simple geometric relations have great practical utility as well because a vast assortment of problems can be analyzed from this point of view.

In this section, we briefly describe a few of the important geometric principles of optimization that are developed in detail in later chapters.

1. The Projection Theorem. This theorem is one of the simplest and nicest results of optimization theory. In ordinary three-dimensional Euclidean space, it states that the shortest line from a point to a plane is furnished by the perpendicular from the point to the plane, as illustrated in Figure 1.2.

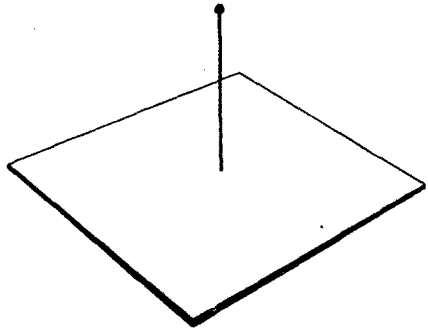

Figure 1.2 The projection theorem

This simple and seemingly innocuous result has direct extensions in spaces of higher dimension and in infinite-dimensional Hilbert space. In the generalized form, this optimization principle forms the basis of all least-squares approximation, control, and estimation procedures.

2. The Hahn-Banach Theorem. Of the many results and concepts in functional analysis, the one theorem dominating the theme of this book and embodying the essence of the simple geometric ideas upon which the theory is built is the Hahn-Banach theorem. The theorem takes several forms. One version extends the projection theorem to problems having nonquadratic objectives. In this manner the simple geometric interpretation is preserved for these more complex problems. Another version of the

Hahn-Banach theorem states (in simplest form) that given a sphere and a point not in the sphere there is a hyperplane separating the point and the sphere. This version of the theorem, together with the associated notions of hyperplanes, is the basis for most of the theory beyond Chapter 5.

*3. **Duality.*** There are several duality principles in optimization theory that relate a problem expressed in terms of vectors in a space to a problem expressed in terms of hyperplanes in the space. This concept of duality is a recurring theme in this book.

Many of these duality principles are based on the geometric relation illustrated in Figure 1.3. The shortest distance from a point to a convex set is equal to the maximum of the distances from the point to a hyperplane separating the point from the convex set. Thus, the original minimization over vectors can be converted to maximization over hyperplanes.

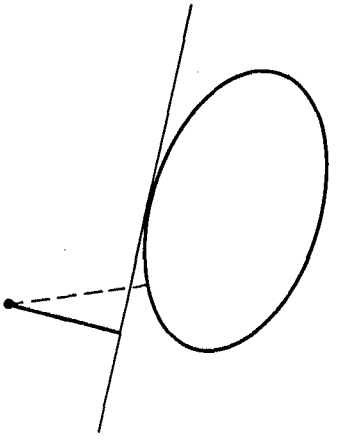

Figure 1.3 Duality

*4. **Differentials.*** Perhaps the most familiar optimization technique is the method of differential calculus—setting the derivative of the objective function equal to zero. The technique is discussed for a single or, perhaps, finite number of variables in the most elementary courses on differential calculus. Its extension to infinite-dimensional spaces is straightforward and, in that form, it can be applied to a variety of interesting optimization problems. Much of the classical theory of the calculus of variations can be viewed as a consequence of this principle.

The geometric interpretation of the technique for one-dimensional problems is obvious. At a maximum or minimum the tangent to the graph of a function is horizontal. In higher dimensions the geometric interpretation is similar: at a maximum or minimum the tangent hyperplane to the

graph is horizontal. Thus, again we are led to observe the fundamental role of hyperplanes in optimization.

1.4 Organization of the Book

Before our discussion of optimization can begin in earnest, certain fundamental concepts and results of linear vector space theory must be introduced. Chapter 2 is devoted to that task. The chapter consists of material that is standard, elementary functional analysis background and is essential for further pursuit of our objectives. Anyone having some familiarity with linear algebra and analysis should have little difficulty with this chapter.

Chapters 3 and 4 are devoted to the projection theorem in Hilbert space and its applications. Chapter 3 develops the general theory, illustrating it with some applications from Fourier approximation and optimal control theory. Chapter 4 deals solely with the applications of the projection theorem to estimation problems including the recursive estimation and prediction of time series as developed by Kalman.

Chapter 5 is devoted to the Hahn-Banach theorem. It is in this chapter that we meet with full force the essential ingredients of the general theory of optimization: hyperplanes, duality, and convexity.

Chapter 6 discusses linear transformations on a vector space and is the last chapter devoted to the elements of linear functional analysis. The concept of duality is pursued in this chapter through the introduction of adjoint transformations and their relation to minimum norm problems. The pseudoinverse of an operator in Hilbert space is discussed.

Chapters 7, 8, and 9 consider general optimization problems in linear spaces. Two basic approaches, the local theory leading to differential conditions and the global theory relying on convexity, are isolated and discussed in a parallel fashion. The techniques in these chapters are a direct outgrowth of the principles of earlier chapters, and geometric visualization is stressed wherever possible. In the course of the development, we treat problems from the calculus of variations, the Fenchel conjugate function theory, Lagrange multipliers, the Kuhn-Tucker theorem, and Pontryagin's maximum principle for optimal control problems.

Finally, Chapter 10 contains an introduction to iterative techniques for the solution of optimization problems. Some techniques in this chapter are quite different than those in previous chapters, but many are based on extensions of the same logic and geometrical considerations found to be so fruitful throughout the book. The methods discussed include successive approximation, Newton's method, steepest descent, conjugate gradients, the primal-dual method, and penalty functions.

2

LINEAR SPACES

2.1 Introduction

The first few sections of this chapter introduce the concept of a vector space and explore the elementary properties resulting from the basic definition. The notions of subspace, linear independence, convexity, and dimension are developed and illustrated by examples. The material is largely review for most readers since it duplicates the first part of standard courses in linear algebra.

The second part of the chapter discusses the basic properties of normed linear spaces. A normed linear space is a vector space having a measure of distance or length defined on it. With the introduction of a norm, it becomes possible to define analytical or topological properties such as convergence and open and closed sets. Therefore, that portion of the chapter introduces and explores these basic concepts which distinguish functional analysis from linear algebra.

VECTOR SPACES

2.2 Definition and Examples

Associated with every vector space is a set of scalars used to define scalar multiplication on the space. In the most abstract setting these scalars are required only to be elements of an algebraic field. However, in this book the scalars are always taken to be either the set of real numbers or of complex numbers. We sometimes distinguish between these possibilities by referring to a vector space as either a real or a complex vector space. In this book, however, the primary emphasis is on real vector spaces and, although occasional reference is made to complex spaces, many results are derived only for real spaces. In case of ambiguity, the reader should assume the space to be real.

Definition. A *vector space* X is a set of elements called vectors together with two operations. The first operation is addition which associates with

any two vectors $x, y \in X$ a vector $x + y \in X$, the sum of x and y. The second operation is scalar multiplication which associates with any vector $x \in X$ and any scalar α a vector αx; the scalar multiple of x by α. The set X and the operations of addition and scalar multiplication are assumed to satisfy the following axioms:

1. $x + y = y + x$. (commutative law)
2. $(x + y) + z = x + (y + z)$. (associative law)
3. There is a null vector θ in X such that $x + \theta = x$ for all x in X.
4. $\alpha(x + y) = \alpha x + \alpha y$. } (distributive laws)
5. $(\alpha + \beta)x = \alpha x + \beta x$. }
6. $(\alpha\beta)x = \alpha(\beta x)$. (associative law)
7. $0x = \theta, \quad 1x = x$.

For convenience the vector $-1x$ is denoted $-x$ and called the negative of the vector x. We have $x + (-x) = (1 - 1)x = 0x = \theta$.

There are several elementary but important properties of vector spaces that follow directly from the axioms listed in the definition. For example the following properties are easily deduced. The details are left to the reader.

Proposition 1. *In any vector space:*

1. $x + y = x + z$ *implies* $y = z$. }
2. $\alpha x = \alpha y$ *and* $\alpha \neq 0$ *imply* $x = y$. } (*cancellation laws*)
3. $\alpha x = \beta x$ *and* $x \neq \theta$ *imply* $\alpha = \beta$. }
4. $(\alpha - \beta)x = \alpha x - \beta x$. } (*distributive laws*)
5. $\alpha(x - y) = \alpha x - \alpha y$. }
6. $\alpha\theta = \theta$.

Some additional properties are given as exercises at the end of the chapter.

Example 1. Perhaps the simplest example of a vector space is the set of real numbers. It is a real vector space with addition defined in the usual way and multiplication by (real) scalars defined as ordinary multiplication. The null vector is the real number zero. The properties of ordinary addition and multiplication of real numbers satisfy the axioms in the definition of a vector space. This vector space is called the one-dimensional real coordinate space or simply the real line. It is denoted by R^1 or simply R.

Example 2. An obvious extension of Example 1 is to n-dimensional real coordinate space. Vectors in the space consist of sequences (n-tuples) of n real numbers so that a typical vector has the form $x = (\xi_1, \xi_2, \ldots, \xi_n)$.

The real number ξ_k is referred to as the k-th component of the vector. Two vectors are equal if their corresponding components are equal. The null vector is defined as $\theta = (0, 0, \ldots, 0)$. If $x = (\xi_1, \xi_2, \ldots, \xi_n)$ and $y = (\eta_1, \eta_2, \ldots, \eta_n)$, the vector $x + y$ is defined as the n-tuple whose k-th component is $\xi_k + \eta_k$. The vector αx, where α is a (real) scalar, is the n-tuple whose k-th component is $\alpha\xi_k$. The axioms in the definition are verified by checking for equality among components. For example, if $x = (\xi_1, \xi_2, \ldots, \xi_n)$, the relation $\xi_k + 0 = \xi_k$ implies $x + \theta = x$.

This space, n-dimensional real coordinate space, is denoted by R^n. The corresponding complex space consisting of n-tuples of complex numbers is denoted by C^n.

At this point we are, strictly speaking, somewhat prematurely introducing the term dimensionality. Later in this chapter the notion of dimension is defined, and it is proved that these spaces are in fact n dimensional.

Example 3. Several interesting vector spaces can be constructed with vectors consisting of infinite sequences of real numbers so that a typical vector has the form $x = (\xi_1, \xi_2, \ldots, \xi_k, \ldots)$ or, equivalently, $x = \{\xi_k\}_{k=1}^{\infty}$. Again addition and multiplication are defined componentwise as in Example 2. The collection of all infinite sequences of real numbers forms a vector space. A sequence $\{\xi_k\}$ is said to be bounded if there is a constant M such that $|\xi_k| < M$ for all k. The collection of all bounded infinite sequences forms a vector space since the sum of two bounded sequences or the scalar multiple of a bounded sequence is again bounded. This space is referred to as the space of bounded real sequences.

Example 4. The collection of all sequences of real numbers having only a finite number of terms not equal to zero is a vector space. (Different members of the space may have different numbers of nonzero components.) This space is called the space of finitely nonzero sequences.

Example 5. The collection of infinite sequences of real numbers which converge to zero is a vector space since the sum of two sequences converging to zero or the scalar multiple of a sequence converging to zero also converges to zero.

Example 6. Consider the interval $[a, b]$ on the real line. The collection of all real-valued continuous functions on this interval forms a vector space. Write $x = y$ if $x(t) = y(t)$ for all $t \in [a, b]$. The null vector θ is the function identically zero on $[a, b]$. If x and y are vectors in the space and α is a (real) scalar, write $(x + y)(t) = x(t) + y(t)$ and $(\alpha x)(t) = \alpha x(t)$. These are obviously continuous functions. This space is referred to as the vector space of real-valued continuous functions on $[a, b]$.

Example 7. The collection of all polynomial functions defined on the interval $[a, b]$ with complex coefficients forms a complex vector space. The null vector, addition, and scalar multiplication are defined as in Example 6. The sum of two polynomials and the scalar multiple of a polynomial are themselves polynomials.

We now consider how a set of vector spaces can be combined to produce a larger one.

Definition. Let X and Y be vector spaces over the same field of scalars. Then the *Cartesian product* of X and Y, denoted $X \times Y$, consists of the collection of ordered pairs (x, y) with $x \in X$, $y \in Y$. Addition and scalar multiplication are defined on $X \times Y$ by $(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$ and $\alpha(x, y) = (\alpha x, \alpha y)$.

That the above definition is consistent with the axioms of a vector space is obvious. The definition is easily generalized to the product of n vector spaces $X_1, X_2, \ldots, X_n$. We write X^n for the product of a vector space with itself n times.

2.3 Subspaces, Linear Combinations, and Linear Varieties

Definition. A nonempty subset M of a vector space X is called a *subspace* of X if every vector of the form $\alpha x + \beta y$ is in M whenever x and y are both in M.

Since a subspace is assumed to be nonempty, it must contain at least one element x. By definition, it must then also contain $0x = \theta$, so every subspace contains the null vector. The simplest subspace is the set consisting of θ alone. In three-dimensional space a plane that passes through the origin is a subspace as is a line through the origin.

The entire space X is itself a subspace of X. A subspace not equal to the entire space is said to be a *proper subspace.*

Any subspace contains sums and scalar multiples of its elements; satisfaction of the seven axioms in the entire space implies that they are satisfied within the subspace. Therefore a subspace is itself a vector space and this observation justifies the terminology.

If X is the space of n-tuples, the set of n-tuples having the first component equal to zero is a subspace of X. The space of convergent infinite sequences is a subspace of the vector space of bounded sequences. The space of continuous functions on $[0, 1]$ which are zero at the point one-half is a subspace of the vector space of continuous functions.

In three-dimensional space, the intersection of two distinct planes

containing the origin is a line containing the origin. This is a special case of the following result.

Proposition 1. *Let M and N be subspaces of a vector space X. Then the intersection, $M \cap N$, of M and N is a subspace of X.*

Proof. $M \cap N$ contains θ since 0 is contained in each of the subspaces M and N. Therefore, $M \cap N$ is nonempty. If x and y are in $M \cap N$, they are in both M and N. For any scalars α, β the vector $\alpha x + \beta y$ is contained in both M and N since M and N are subspaces. Therefore, $\alpha x + \beta y$ is contained in the intersection $M \cap N$. ∎

The union of two subspaces in a vector space is not necessarily a subspace. In the plane, for example, the union of two (noncolinear) lines through the origin does not contain arbitrary sums of its elements. However, two subspaces may be combined to form a larger subspace by introducing the notion of the sum of two sets.

Definition. The *sum* of two subsets S and T in a vector space, denoted $S + T$, consists of all vectors of the form $s + t$ where $s \in S$ and $t \in T$.

The sum of two sets is illustrated in Figure 2.1.

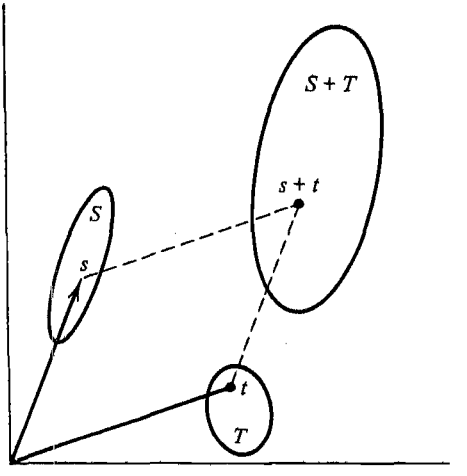

Figure 2.1 The sum of two sets

Proposition 2. *Let M and N be subspaces of a vector space X. Then their sum, $M + N$, is a subspace of X.*

Proof. Clearly $M + N$ contains θ. Suppose x and y are vectors in $M + N$. There are vectors m_1, m_2 in M and vectors n_1, n_2 in N such that

$x = m_1 + n_1, y = m_2 + n_2$. For any scalars α, β; $\alpha x + \beta y = (\alpha m_1 + \beta m_2) + (\alpha n_1 + \beta n_2)$. Therefore, $\alpha x + \beta y$ can be expressed as the sum of a vector in the subspace M and a vector in the subspace N. ▮

In two-dimensional Euclidean space the sum of two noncolinear lines through the origin is the entire space. The set of even continuous functions and the set of odd continuous functions are subspaces of the space X of continuous functions on the real line. Their sum is the whole space X since any continuous function can be expressed as a sum of an even and an odd continuous function.

Definition. A *linear combination* of the vectors $x_1, x_2, \ldots, x_n$ in a vector space is a sum of the form $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$.

Actually, addition has been defined previously only for a sum of two vectors. To form a sum consisting of n elements, the sum must be performed two at a time. It follows from the axioms, however, that analogous to the corresponding operations with real numbers, the result is independent of the order of summation. There is thus no ambiguity in the simplified notation.

It is apparent that a linear combination of vectors from a subspace is also in the subspace. Conversely, linear combinations can be used to construct a subspace from an arbitrary subset in a vector space.

Definition. Suppose S is a subset of a vector space X. The set $[S]$, called the *subspace generated by* S, consists of all vectors in X which are linear combinations of vectors in S.

The verification that $[S]$ is a subspace in X follows from the obvious fact that a linear combination of linear combinations is also a linear combination.

There is an interesting characterization of the subspace $[S]$. The set S is, in general, wholly contained in a number of subspaces. Of these, the subspace $[S]$ is the smallest in the sense that if M is a subspace containing S, then M contains $[S]$. This statement is proved by noting that if the subspace M contains S, it must contain all linear combinations from S.

In three-dimensional space the subspace generated by a two-dimensional circle centered at the origin is a plane. The subspace generated by a plane not passing through the origin is the whole space. A subspace is a generalization of our intuitive notion of a plane or line through the origin. The translation of a subspace, therefore, is a generalization of an arbitrary plane or line.

Definition. The translation of a subspace is said to be a *linear variety*.

A linear variety[1] V can be written as $V = x_0 + M$ where M is a subspace. In this representation the subspace M is unique, but any vector in V can serve as x_0. This is illustrated in Figure 2.2.

Given a subset S, we can construct the smallest linear variety containing S.

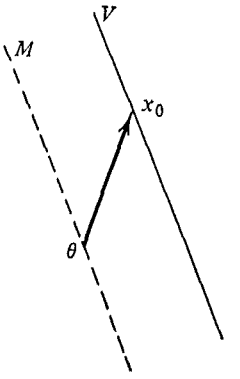

Figure 2.2 A linear variety

Definition. Let S be a nonempty subset of a vector space X. The *linear variety generated by* S, denoted $v(S)$ is defined as the intersection of all linear varieties in X that contain S.

We leave it to the reader to justify the above definition by showing that $v(S)$ is indeed a linear variety.

2.4 Convexity and Cones

We come now to the topic that is responsible for a surprising number of the results in this book and which generalizes many of the useful properties of subspaces and linear varieties.

Definition. A set K in a linear vector space is said to be *convex* if, given $x_1, x_2 \in K$, all points of the form $\alpha x_1 + (1 - \alpha)x_2$ with $0 \leq \alpha \leq 1$ are in K.

This definition merely says that given two points in a convex set, the line segment between them is also in the set. Examples are shown in Figure 2.3 Note in particular that subspaces and linear varieties are convex. The empty set is considered convex.

[1] Other names for a *linear variety* include: flat, affine subspace, and linear manifold. The term linear manifold, although commonly used, is reserved by many authors as an alternative term for subspace.

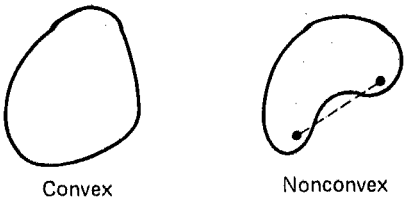

Figure 2.3 Convex and nonconvex sets

The following relations for convex sets are elementary but important. Their proofs are left to the reader.

Proposition 1. *Let K and G be convex sets in a vector space. Then*

1. $\alpha K = \{x: x = \alpha k, k \in K\}$ *is convex for any scalar* α.
2. $K + G$ *is convex.*

We also have the following elementary property.

Proposition 2. *Let $\mathscr{C}$ be an arbitrary collection of convex sets. Then* $\bigcap_{K \in \mathscr{C}} K$ *is convex.*

Proof. Let $C = \bigcap_{K \in \mathscr{C}} K$. If C is empty, the lemma is trivially proved. Assume that $x_1, x_2 \in C$ and select α, $0 \leq \alpha \leq 1$. Then $x_1, x_2 \in K$ for all $K \in \mathscr{C}$, and since K is convex, $\alpha x_1 + (1 - \alpha)x_2 \in K$ for all $K \in \mathscr{C}$. Thus $\alpha x_1 + (1 - \alpha)x_2 \in C$ and C is convex. ▮

Definition. Let S be an arbitrary set in a linear vector space. The *convex cover* or *convex hull*, denoted $\text{co}(S)$ is the smallest convex set containing S. In other words, $\text{co}(S)$ is the intersection of all convex sets containing S.

Note that the justification of this definition rests with Proposition 2 since it guarantees the existence of a smallest convex set containing S. Some examples of a set and its convex hull are illustrated in Figure 2.4.

Definition. A set C in a linear vector space is said to be a *cone with vertex at the origin* if $x \in C$ implies that $\alpha x \in C$ for all $\alpha \geq 0$.

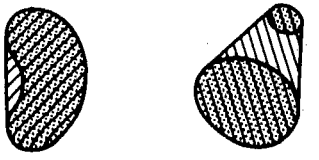

Figure 2.4 Convex hulls

Several cases are shown in Figure 2.5. A *cone with vertex p* is defined as a translation $p + C$ of a cone C with vertex at the origin. If the vertex of a cone is not explicitly mentioned, it is assumed to be the origin. A *convex cone* is, of course, defined as a set which is both convex and a cone. Of the cones in Figure 2.5, only (*b*) is a convex cone. Cones again generalize the concepts of subspace and linear variety since both of these are convex cones.

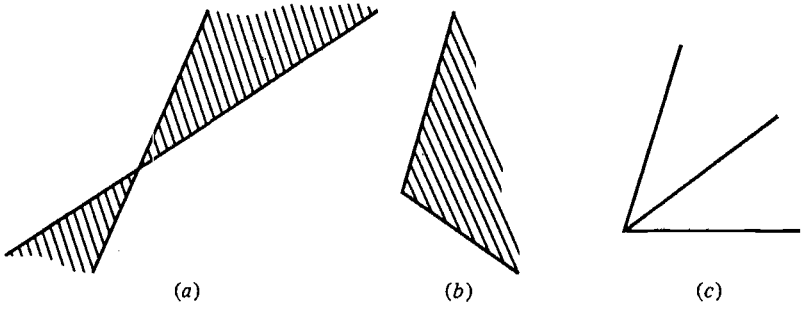

Figure 2.5 Cones

Convex cones usually arise in connection with the definition of positive vectors in a vector space. For instance, in n-dimensional real coordinate space, the set

$$P = \{x : x = \{\xi_1, \xi_2, \ldots, \xi_n\}, \xi_i \geq 0 \quad \text{all} \quad i\},$$

defining the positive orthant, is a convex cone. Likewise, the set of all nonnegative continuous functions is a convex cone in the vector space of continuous functions.

2.5 Linear Independence and Dimension

In this section we first introduce the concept of linear independence, which is important for any general study of vector space, and then review the essential distinguishing features of finite-dimensional space: basis and dimension.

Definition. A vector x is said to be *linearly dependent* upon a set S of vectors if x can be expressed as a linear combination of vectors from S. Equivalently, x is linearly dependent upon S if x is in $[S]$, the subspace generated by S. Conversely, the vector x is said to be *linearly independent* of the set S if it is not linearly dependent on S; a set of vectors is said to be a *linearly independent set* if each vector in the set is linearly independent of the remainder of the set.

Thus, two vectors are linearly independent if they do not lie on a common line through the origin, three vectors are linearly independent if they do not lie in a plane through the origin, etc. It follows from our definition that the vector θ is dependent on any given vector x since $\theta = 0x$. Also, by convention, the set consisting of θ alone is understood to be a dependent set. On the other hand, a set consisting of a single nonzero vector is an independent set. With these conventions the following major test for linear independence is applicable even to trivial sets consisting of a single vector.

Theorem 1. *A necessary and sufficient condition for the set of vectors* $x_1, x_2, \ldots, x_n$ *to be linearly independent is that the expression* $\sum_{k=1}^{n} \alpha_k x_k = \theta$ *implies* $\alpha_k = 0$ *for all* $k = 1, 2, 3, \ldots, n$.

Proof. To prove the necessity of the condition, assume that $\sum_{k=1}^{n} \alpha_k x_k = \theta$, and that for some index r, $\alpha_r \neq 0$. Since the coefficient of x_r is nonzero, the original relation may be rearranged to produce $x_r = \sum_{k \neq r} (-\alpha_k/\alpha_r)x_k$ which shows x_r to be linearly dependent on the remaining vectors.

To prove the sufficiency of the condition, note that linear dependence among the vectors implies that one vector, say x_r, can be expressed as a linear combination of the others, $x_r = \sum_{k \neq r} \alpha_k x_k$. Rearrangement gives $\sum_{k \neq r} \alpha_k x_k - x_r = \theta$ which is the desired relation. ∎

An important consequence of this theorem is that a vector expressed as a linear combination of linearly independent vectors can be so expressed in only one way.

Corollary 1. *If* $x_1, x_2, \ldots, x_n$ *are linearly independent vectors, and if* $\sum_{k=1}^{n} \alpha_k x_k = \sum_{k=1}^{n} \beta_k x_k$, *then* $\alpha_k = \beta_k$ *for all* $k = 1, 2, \ldots, n$.

Proof. If $\sum_{k=1}^{n} \alpha_k x_k = \sum_{k=1}^{n} \beta_k x_k$, then $\sum_{k=1}^{n} (\alpha_k - \beta_k)x_k = \theta$ and $\alpha_k - \beta_k = 0$ according to Theorem 1. ∎

We turn now from the general notion of linear independence to the special topic of finite dimensionality. Consider a set S of linearly independent vectors in a vector space. These vectors may be used to generate $[S]$, a certain subspace of the vector space. If the subspace $[S]$ is actually the entire vector space, every vector in the space is expressible as a linear combination of vectors from the original set S. Furthermore, according to the corollary to Theorem 1, the expression is unique.

Definition. A finite set S of linearly independent vectors is said to be a *basis* for the space X if S generates X. A vector space having a finite basis is said to be *finite dimensional*. All other vector spaces are said to be *infinite dimensional*.

Usually, we characterize a finite-dimensional space by the number of elements in a basis. Thus, a space with a basis consisting of n elements is referred to as n-dimensional space. This practice would be undesirable on grounds of ambiguity if the number of elements in a basis for a space were not unique. The following theorem is, for this reason, a fundamental result in the study of finite-dimensional spaces.

Theorem 2. *Any two bases for a finite-dimensional vector space contain the same number of elements.*

Proof. Suppose that $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ are bases for a vector space X. Suppose also that $m \geq n$. We shall substitute y vectors for x vectors one by one in the first basis until all the x vectors are replaced.

Since the x_i's form a basis, the vector y_1 may be expressed as a linear combination of them, say, $y_1 = \sum_{i=1}^{n} \alpha_i x_i$. Since $y_1 \neq \theta$, at least one of the scalars α_i must be nonzero. Rearranging the x_i's if necessary, it may be assumed that $\alpha_1 \neq 0$. Then x_1 may be expressed as a linear combination of $y_1, x_2, x_3, \ldots, x_n$ by the formula

$$x_1 = \alpha_1^{-1} y_1 + \sum_{i=2}^{n} \alpha_1^{-1} \alpha_i x_i.$$

The set $y_1, x_2, \ldots, x_n$ generates X since any linear combination of the original x_i's becomes an equivalent linear combination of this new set when x_1 is replaced according to the above formula.

Suppose now that $k-1$ of the vectors x_i have been replaced by the first $k-1$ y_i's. The vector y_k can be expressed as a linear combination of $y_1, y_2, \ldots, y_{k-1}, x_k, \ldots, x_n$, say,

$$y_k = \sum_{i=1}^{k-1} \alpha_i y_i + \sum_{i=k}^{n} \beta_i x_i.$$

Since the vectors $y_1, y_2, \ldots, y_k$ are linearly independent, not all the β_i's can be zero. Rearranging $x_k, x_{k+1}, \ldots, x_n$ if necessary, it may be assumed that $\beta_k \neq 0$. Then x_k can be solved for as a linear combination of $y_1, y_2, \ldots, y_k, x_{k+1}, \ldots, x_n$, and this new set of vectors generates X.

By induction on k then, we can replace all n of the x_i's by y_i's, forming a generating set at each step. This implies that the independent vectors $y_1, y_2, \ldots, y_n$ generate X and hence form a basis for X. Therefore, by the assumption of linear independence of $\{y_1, y_2, \ldots, y_m\}$, we must have $n = m$. ∎

Finite-dimensional spaces are somewhat simpler to analyze than infinite-dimensional spaces. Fewer definitions are required, fewer pathological cases arise, and our native intuition is contradicted in fewer instances

in finite-dimensional spaces. Furthermore, there are theorems and concepts in finite-dimensional space which have no direct counterpart in infinite-dimensional spaces. Occasionally, the availability of these special characteristics of finite dimensionality is essential to obtaining a solution to a particular problem. It is more usual, however, that results first derived for finite-dimensional spaces do have direct analogs in more general spaces. In these cases, verification of the corresponding result in infinite-dimensional space often enhances our understanding by indicating precisely which properties of the space are responsible for the result. We constantly endeavor to stress the similarities between infinite- and finite-dimensional spaces rather than the few minor differences.

NORMED LINEAR SPACES

2.6 Definition and Examples

The vector spaces of particular interest in both abstract analysis and applications have a good deal more structure than that implied solely by the seven principal axioms. The vector space axioms only describe algebraic properties of the elements of the space: addition, scalar multiplication, and combinations of these. What are missing are the topological concepts such as openness, closure, convergence, and completeness. These concepts can be provided by the introduction of a measure of distance in a space.

Definition. A *normed linear vector space* is a vector space X on which there is defined a real-valued function which maps each element x in X into a real number $\|x\|$ called the norm of x. The norm satisfies the following axioms:

1. $\|x\| \geq 0$ for all $x \in X$, $\|x\| = 0$ if and only if $x = \theta$.
2. $\|x + y\| \leq \|x\| + \|y\|$ for each $x, y \in X$. (triangle inequality)
3. $\|\alpha x\| = |\alpha| \cdot \|x\|$ for all scalars α and each $x \in X$.

The norm is clearly an abstraction of our usual concept of length. The following useful inequality is a direct consequence of the triangle inequality.

Lemma 1. *In a normed linear space* $\|x\| - \|y\| \leq \|x - y\|$ *for any two vectors* x, y.

Proof.

$$\|x\| - \|y\| = \|x - y + y\| - \|y\| \leq \|x - y\| + \|y\| - \|y\| = \|x - y\|. \blacksquare$$

By introduction of a suitable norm, many of our earlier examples of vector spaces can be converted to normed spaces.

Example 1. The normed linear space $C[a, b]$ consists of continuous functions on the real interval $[a, b]$ together with the norm $\|x\| = \max_{a \le t \le b} |x(t)|$. This space was considered as a vector space in Section 2.2. We now verify that the proposed norm satisfies the three required axioms. Obviously, $\|x\| \ge 0$ and is zero only for the function which is identically zero. The triangle inequality follows from the relation

$$\max |x(t) + y(t)| \le \max [|x(t)| + |y(t)|] \le \max |x(t)| + \max |y(t)|.$$

Finally, the third axiom follows from the relation

$$\max |\alpha x(t)| = \max |\alpha| \, |x(t)| = |\alpha| \max |x(t)|.$$

Example 2. The normed linear space $D[a, b]$ consists of all functions on the interval $[a, b]$ which are continuous and have continuous derivatives on $[a, b]$. The norm on the space $D[a, b]$ is defined as

$$\|x\| = \max_{a \le t \le b} |x(t)| + \max_{a \le t \le b} |\dot{x}(t)|.$$

We leave it to the reader to verify that $D[a, b]$ is a normed linear space.

Example 3. The space of finitely nonzero sequences together with the norm equal to the sum of the absolute values of the nonzero components is a normed linear space. Thus the element $x = \{\xi_1, \xi_2, \ldots, \xi_n, 0, 0, \ldots\}$ has its norm defined as $\|x\| = \sum_{i=1}^{n} |\xi_i|$. We may easily verify the three required properties by inspection.

Example 6. The space of continuous functions on the interval $[a, b]$ becomes a normed space with the norm of a function x defined as $\|x\| = \int_a^b |x(t)| \, dt$. This is a different normed space than $C[a, b]$.

Example 5. Euclidean n-space, denoted E^n, consists of n-tuples with the norm of an element $x = \{\xi_1, \xi_2, \ldots, \xi_n\}$ defined as $\|x\| = \left(\sum_{i=1}^{n} |\xi_i|^2\right)^{1/2}$. This definition obviously satisfies the first and third axioms for norms. The triangle inequality for this norm is a well-known result from finite-dimensional vector spaces and is a special case of the Minkowski inequality discussed in Section 2.10. The space E^n can be chosen as a real or complex space by considering real or complex n-tuples. We employ the same notation E^n for both because it is generally apparent from context which is meant.

Example 6. We consider now the space $BV[a, b]$ consisting of functions of bounded variation on the interval $[a, b]$. By a partition of the interval $[a, b]$, we mean a finite set of points $t_i \in [a, b]$, $i = 0, 1, 2, \ldots, n$, such that $a = t_0 < t_1 < t_2 < \cdots < t_n = b$. A function x defined on $[a, b]$ is said to

be of bounded variation if there is a constant K so that for any partition of $[a, b]$

$$\sum_{i=1}^{n} |x(t_i) - x(t_{i-1})| \leq K.$$

The total variation of x, denoted T.V.(x), is then defined as

$$\text{T.V.}(x) = \sup \sum_{i=1}^{n} |x(t_i) - x(t_{i-1})|$$

where the supremum is taken with respect to all partitions of $[a, b]$. A convenient and suggestive notation for the total variation is

$$\text{T.V.}(x) = \int_a^b |dx(t)|.$$

The total variation of a constant function is zero and the total variation of a monotonic function is the absolute value of the difference between the function values at the end points a and b.

The space $BV[a, b]$ is defined as the space of all functions of bounded variation on $[a, b]$ together with the norm defined as

$$\|x\| = |x(a)| + \text{T.V.}(x).$$

2.7 Open and Closed Sets

We come now to the concepts that are fundamental to the study of topological properties.

Definition. Let P be a subset of a normed space X. The point $p \in P$ is said to be an *interior point* of P if there is an $\varepsilon > 0$ such that all vectors x satisfying $\|x - p\| < \varepsilon$ are also members of P. The collection of all interior points of P is called the *interior* of P and is denoted $\mathring{P}$.

We introduce the notation $S(x, \varepsilon)$ for the (open) sphere centered at x with radius ε; that is, $S(x, \varepsilon) = \{y : \|x - y\| < \varepsilon\}$. Thus, according to the above definition, a point x is an interior point of P if there is a sphere $S(x, \varepsilon)$ centered at x and contained in P. A set may have an empty interior as, for example, a set consisting of a single point or a line in E^2.

Definition. A set P is said to be *open* if $P = \mathring{P}$.

The empty set is open since its interior is also empty. The entire space is an open set. The unit sphere consisting of all vectors x with $\|x\| < 1$ is an open set. We leave it to the reader to verify that $\mathring{P}$ is open.

Definition. A point $x \in X$ is said to be a *closure point* of a set P if, given $\varepsilon > 0$, there is a point $p \in P$ satisfying $\|x - p\| < \varepsilon$. The collection of all closure points of P is called the *closure* of P and is denoted $\bar{P}$.

In other words, a point x is a closure point of P if every sphere centered at x contains a point of P. It is clear that $P \subset \bar{P}$.

Definition. A set P is said to be *closed* if $P = \bar{P}$.

The empty set and the whole space X are closed as well as open. The unit sphere consisting of all points x with $\|x\| \leq 1$ is a closed set. A single point is a closed set. It is clear that $\bar{\bar{P}} = \bar{P}$.

Proposition 1. *The complement of an open set is closed and the complement of a closed set is open.*

Proof. Let P be an open set and $\tilde{P} = \{x : x \notin P\}$ its complement. A point x in P is not a closure point of $\tilde{P}$ since there is a sphere about x disjoint from $\tilde{P}$. Thus $\tilde{P}$ contains all its closure points and is therefore closed.

Let S be a closed set. If $x \in \tilde{S}$, then x is not a closure point of S and, hence, there is a sphere about x which is disjoint from $\tilde{S}$. Therefore x is an interior point of $\tilde{S}$. We conclude that $\tilde{S}$ is open. ∎

The proofs of the following two complementary results are left to the reader.

Proposition 2. *The intersection of a finite number of open sets is open; the union of an arbitrary collection of open sets is open.*

Proposition 3. *The union of a finite number of closed sets is closed; the intersection of an arbitrary collection of closed sets is closed.*

We now have two topological operations, taking closures and taking interiors, that can be applied to sets in normed space. It is natural to investigate the effect of these operations on convexity, the fundamental algebraic concept of vector space.

Proposition 4. *Let C be a convex set in a normed space. Then $\bar{C}$ and $\mathring{C}$ are convex.*

Proof. If $\bar{C}$ is empty, it is convex. Suppose x_0, y_0 are points in $\bar{C}$. Fix α, $0 < \alpha < 1$. We must show that $z_0 = \alpha x_0 + (1 - \alpha)y_0 \in \bar{C}$. Given $\varepsilon > 0$, let x, y be selected from C such that $\|x - x_0\| < \varepsilon$, $\|y - y_0\| < \varepsilon$. Then $\|\alpha x + (1 - \alpha)y - \alpha x_0 - (1 - \alpha)y_0\| \leq \varepsilon$ and, hence, z_0 is within a distance ε of $z = \alpha x + (1 - \alpha)y$ which is in C. Since ε is arbitrary, it follows that z_0 is a closure point of C.

If $\mathring{C}$ is empty, it is convex. Suppose $x_0, y_0 \in \mathring{C}$ and fix α, $0 < \alpha < 1$. We must show that $z_0 = \alpha x_0 + (1 - \alpha)y_0 \in \mathring{C}$. Since $x_0, y_0 \in \mathring{C}$, there is an $\varepsilon > 0$ such that the open spheres $S(x_0, \varepsilon)$, $S(y_0, \varepsilon)$ are contained in C. It follows that all points of the form $z_0 + w$ with $\|w\| < \varepsilon$ are in C since $z_0 + w = \alpha(x_0 + w) + (1 - \alpha)(y_0 + w)$. Thus, z_0 is an interior point of C. ∎

Likewise, it can be shown that taking closure preserves subspaces, linear varieties, and cones.

Finally, we remark that all of the topological concepts discussed above can be defined relative to a given linear variety. Suppose that P is a set contained in a linear variety V. We say that $p \in P$ is an *interior point of P relative to V* if there is an $\varepsilon > 0$ such that all vectors $x \in V$ satisfying $\|x - p\| < \varepsilon$ are also members of P. The set P is said to be *open relative to V* if every point in P is an interior point of P relative to V.

In case V is taken as the closed linear variety generated by P, i.e., the intersection of all closed linear varieties containing P, then x is simply referred to as a *relative interior point* of P if it is an interior point of P relative to the variety V. Similar meaning is given to *relatively closed*, etc.

2.8 Convergence

In order to prove the existence of a vector satisfying a desired property, it is common to establish an appropriate sequence of vectors converging to a limit. In many cases the limit can be shown to satisfy the required property. It is for this reason that the concept of convergence plays an important role in analysis.

Definition. In a normed linear space an infinite sequence of vectors $\{x_n\}$ is said to *converge* to a vector x if the sequence $\{\|x - x_n\|\}$ of real numbers converges to zero. In this case, we write $x_n \to x$.

If $x_n \to x$, it follows that $\|x_n\| \to \|x\|$ because, according to Lemma 1, Section 2.6, we have both $\|x_n\| - \|x\| \leq \|x_n - x\|$ and $\|x\| - \|x_n\| \leq \|x_n - x\|$ which implies that $\big|\, \|x_n\| - \|x\| \,\big| \leq \|x_n - x\| \to 0$.

In the space E^n of n-tuples, a sequence converges if and only if each component converges; however, in other spaces convergence is not always easy to characterize. In the space of finitely nonzero sequences, define the vectors $e_i = \{0, 0, \ldots, 1, 0, \ldots\}$, the i-th vector having each of its components zero except the i-th which is 1. The sequence of vectors $\{e_i\}$ (which is now a sequence of sequences) converges to zero componentwise, but the sequence does not converge to the null vector since $\|e_i\| = 1$ for all i.

An important observation is stated in the following proposition.

Proposition 1. *If a sequence converges, its limit is unique.*

Proof. Suppose $x_n \to x$ and $x_n \to y$. Then

$$\|x - y\| = \|x - x_n + x_n - y\| \leq \|x - x_n\| + \|x_n - y\| \to 0.$$

Thus, $x = y$. ∎

Another way to state the definition of convergence is in terms of spheres. A sequence $\{x_n\}$ converges to x if and only if given $\varepsilon > 0$; the sphere $S(x, \varepsilon)$ contains x_n for all n greater than some number N.

The definition of convergence can be used to characterize closed sets and provides a useful alternative to the original definition of closed sets.

Proposition 2. *A set F is closed if and only if every convergent sequence with elements in F has its limit in F.*

Proof. The limit of a sequence from F is obviously a closure point of F and, therefore, must be contained in F if F is closed. Suppose now that F is not closed. Then there is a closure point x of F that is not in F. In each of the spheres $S(x, 1/n)$ we may select a point $x_n \in F$ since x is a closure point. The sequence $\{x_n\}$ generated in this way converges to $x \notin F$. ∎

2.9 Transformations and Continuity

The objects that make the study of linear spaces interesting and useful are transformations.

Definition. Let X and Y be linear vector spaces and let D be a subset of X. A rule which associates with every element $x \in D$ an element $y \in Y$ is said to be a *transformation* from X to Y with *domain* D. If y corresponds to x under T, we write $y = T(x)$.

Transformations on vector spaces become increasingly more important as we progress through this book; they are treated in some detail beginning with Chapter 6. It is the purpose of this section to introduce some common terminology that is convenient for describing the simple transformations encountered in the early chapters.

If a specific domain is not explicitly mentioned when discussing a transformation on a vector space X, it is understood that the domain is X itself. If for every $y \in Y$ there is at most one $x \in D$ for which $T(x) = y$, the transformation T is said to be *one-to-one*. If for every $y \in Y$ there is at least one $x \in D$ for which $T(x) = y$, T is said to be *onto* or, more precisely, to map D onto Y. This terminology, as the notion of a transformation itself, is of course an extension of the familiar notion of ordinary functions. A transformation is simply a function defined on one vector space X while

taking values in another vector space Y. A special case of this situation is that in which the space Y is taken to be the real line.

Definition. A transformation from a vector space X into the space of real (or complex) scalars is said to be a *functional* on X.

In order to distinguish functionals from more general transformations, they are usually denoted by lower case letters such as f and g. Hence, $f(x)$ denotes the scalar that f associates with the vector $x \in X$.

On a normed space, $f(x) = \|x\|$ is an example of a functional. On the space $C[0, 1]$, examples of functionals are $f_1(x) = x(\frac{1}{2})$, $f_2(x) = \int_0^1 x(t)\,dt$, $f_3(x) = \max_{0 \le t \le 1} x^4(t)$, etc. Real-valued functionals are of direct interest to optimization theory, since optimization consists of selecting a vector in minimize (or maximize) a prescribed functional.

Definition. A transformation T mapping a vector space X into a vector space Y is said to be *linear* if for every $x_1, x_2 \in X$ and all scalars α_1, α_2 we have $T(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 T(x_1) + \alpha_2 T(x_2)$.

The most familiar example of a linear transformation is supplied by a rectangular $m \times n$ matrix mapping elements of R^n into R^m. An example of a linear transformation mapping $X = C[a, b]$ into X is the integral operator $T(x) = \int_a^b k(t, \tau)x(\tau)\,d\tau$ where $k(t, \tau)$ is a function continuous on the square $a \le t \le b$; $a \le \tau \le b$.

Up to this point we have considered transformations mapping one abstract space into another. If these spaces happen to be normed, it is possible to define the notion of continuity.

Definition. A transformation T mapping a normed space X into a normed space Y is *continuous* at $x_0 \in X$ if for every $\varepsilon > 0$ there is a $\delta > 0$ such that $\|x - x_0\| < \delta$ implies that $\|T(x) - T(x_0)\| < \varepsilon$.

Note that continuity depends on the norm in both the spaces X and Y. If T is continuous at each point $x_0 \in X$, we say that T is *continuous everywhere* or, more simply, that T is *continuous.*

The following characterization of continuity is useful in many proofs involving continuous transformations.

Proposition 1. *A transformation T mapping a normed space X into a normed space Y is continuous at the point $x_0 \in X$ if and only if $x_n \to x_0$ implies $T(x_n) \to T(x_0)$.*

Proof. The "if" portion of the statement is obvious; thus we need only proof the "only if" portion. Let $\{x_n\}$ be a sequence such that $x_n \to x_0$, $T(x_n) \not\to T(x_0)$. Then, for some $\varepsilon > 0$ and every N there is an $n > N$ such

that $\|T(x_n) - T(x_0)\| \geq \varepsilon$. Since $x_n \to x_0$, this implies that for every $\delta > 0$ there is a point x_n with $\|x_n - x_0\| < \delta$ and $\|T(x_n) - T(x)\| > \varepsilon$. This proves the " only if" portion by contraposition. ▮

*2.10 The l_p and L_p Spaces

In this section we discuss some classical normed spaces that are useful throughout the book.

Definition. Let p be a real number $1 \leq p < \infty$.The space l_p consists of all sequences of scalars $\{\xi_1, \xi_2, \ldots\}$ for which

$$\sum_{i=1}^{\infty} |\xi_i|^p < \infty.$$

The norm of an element $x = \{\xi_i\}$ in l_p is defined as

$$\|x\|_p = \left(\sum_{i=1}^{\infty} |\xi_i|^p \right)^{1/p}.$$

The space l_∞ consists of bounded sequences. The norm of an element $x = \{\xi_i\}$ in l_∞ is defined as

$$\|x\|_\infty = \sup_i |\xi_i|.$$

It is obvious that the norm on l_p satisfies $\|\alpha x\| = |\alpha| \, \|x\|$ and that $\|x\| > 0$ for each $x \neq \theta$. In this section, we establish two inequalities concerning the l_p norms, the second of which gives the triangle inequality for these l_p norms. Therefore, the l_p norm indeed satisfies the three axioms required of a general norm. Incidentally, it follows from these properties of the norm that l_p is in fact a linear vector space because, if $x = \{\xi_i\}$, $y = \{\eta_i\}$ are vectors in l_p, then for any scalars α, β, we have $\|\alpha x + \beta y\| \leq |\alpha| \, \|x\| + |\beta| \, \|y\| < \infty$ so that $\alpha x + \beta y$ is a vector in l_p. Since l_p is a vector space and the norm satisfies the three required axioms, we may justifiably refer to l_p as a normed linear vector space.

The following two theorems, although of fundamental importance for a study of the l_p spaces, are somewhat tangential to our main purpose. The reader will lose little by simply scanning the proofs.

Theorem 1. (*The Hölder Inequality*) *If p and q are positive numbers $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, such that $1/p + 1/q = 1$ and if $x = \{\xi_1, \xi_2, \ldots\} \in l_p$ $y = \{\eta_1, \eta_2, \ldots\} \in l_q$, then*

$$\sum_{i=1}^{\infty} |\xi_i \eta_i| \leq \|x\|_p \cdot \|y\|_q.$$

Equality holds if and only if

$$\left(\frac{|\xi_i|}{\|x\|_p}\right)^{1/q} = \left(\frac{|\eta_i|}{\|y\|_q}\right)^{1/p}$$

for each i.

Proof. The cases $p = 1, \infty$; $q = \infty, 1$ are straightforward and are left to the reader. Therefore, it is assumed that $1 < p < \infty$, $1 < q < \infty$. We first prove the auxiliary inequality: For $a \geq 0$, $b \geq 0$, and $0 < \lambda < 1$, we have

$$a^\lambda b^{(1-\lambda)} \leq \lambda a + (1 - \lambda)b$$

with equality if and only if $a = b$.

For this purpose, consider the function

$$f(t) = t^\lambda - \lambda t + \lambda - 1$$

defined for $t \geq 0$. Then $f'(t) = \lambda(t^{\lambda-1} - 1)$. Since $0 < \lambda < 1$, we have $f'(t) > 0$ for $0 < t < 1$ and $f'(t) < 0$ for $t > 1$. It follows that for $t \geq 0$, $f(t) \leq f(1) = 0$ with equality only for $t = 1$. Hence

$$t^\lambda \leq \lambda t + 1 - \lambda$$

with equality only for $t = 1$. If $b \neq 0$, the substitution $t = a/b$ gives the desired inequality, while for $b = 0$ the inequality is trivial.

Applying this inequality to the numbers

$$a = \left(\frac{|\xi_i|}{\|x\|_p}\right)^{p}, \qquad b = \left(\frac{|\eta_i|}{\|y\|_q}\right)^{q} \qquad \text{with } \lambda = \frac{1}{p}, \quad 1 - \lambda = \frac{1}{q},$$

we obtain for each i

$$\frac{|\xi_i \eta_i|}{\|x\|_p \|y\|_q} \leq \frac{1}{p}\left(\frac{|\xi_i|}{\|x\|_p}\right)^{p} + \frac{1}{q}\left(\frac{|\eta_i|}{\|y\|_q}\right)^{q}.$$

Summing this inequality over i, we obtain the Hölder inequality

$$\frac{\sum_{i=1}^{\infty} |\xi_i \eta_i|}{\|x\|_p \|y\|_q} \leq \frac{1}{p} + \frac{1}{q} = 1.$$

The conditions for equality follow directly from the required condition $a = b$ in the auxiliary inequality. ∎

The special case $p = 2$, $q = 2$ is of major importance. In this case, Hölder's inequality becomes the well-known Cauchy-Schwarz inequality for sequences:

$$\sum_{i=1}^{\infty} |\xi_i \eta_i| \leq \left(\sum_{i=1}^{\infty} |\xi_i|^2\right)^{1/2} \left(\sum_{i=1}^{\infty} |\eta_i|^2\right)^{1/2}$$

Using the Hölder inequality, we can establish the triangle inequality for l_p norms.

Theorem 2. (*The Minkowski Inequality*) *If x and y are in l_p, $1 \le p \le \infty$, then so is $x + y$, and $\|x + y\|_p \le \|x\|_p + \|y\|_p$. For $1 < p < \infty$, equality holds if and only if $k_1 x = k_2 y$ for some positive constants k_1 and k_2.*

Proof. The cases $p = 1$ and $p = \infty$ are straightforward. Therefore, it is assumed that $1 < p < \infty$. We first consider finite sums. Clearly we may write

$$\sum_{i=1}^{n} |\xi_i + \eta_i|^p \le \sum_{i=1}^{n} |\xi_i + \eta_i|^{p-1} |\xi_i| + \sum_{i=1}^{n} |\xi_i + \eta_i|^{p-1} |\eta_i|.$$

Applying Hölder's inequality to each summation on the right, we obtain

$$\sum_{i=1}^{n} |\xi_i + \eta_i|^p \le \left(\sum_{i=1}^{n} |\xi_i + \eta_i|^{(p-1)q} \right)^{1/q} \left[\left(\sum_{i=1}^{n} |\xi_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |\eta_i|^p \right)^{1/p} \right]$$

$$= \left(\sum_{i=1}^{n} |\xi_i + \eta_i|^p \right)^{1/q} \left[\left(\sum_{i=1}^{n} |\xi_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |\eta_i|^p \right)^{1/p} \right].$$

Dividing both sides of this inequality by $(\sum_{i=1}^{n} |\xi + \eta_i|^p)^{1/q}$ and taking account of $1 - 1/q = 1/p$, we find

$$\left(\sum_{i=1}^{n} |\xi_i + \eta_i|^p \right)^{1/p} \le \left(\sum_{i=1}^{n} |\xi_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |\eta_i|^p \right)^{1/p}.$$

Letting $n \to \infty$ on the right side of this inequality can only increase its value, so

$$\left(\sum_{i=1}^{n} |\xi_i + \eta_i|^p \right)^{1/p} \le \|x\|_p + \|y\|_p$$

for each n. Therefore, letting $n \to \infty$ on the left side produces $\|x + y\|_p \le \|x\|_p + \|y\|_p$.

The conditions for equality follow from the conditions for equality in the Hölder inequality and are left to the reader. ∎

The L_p and R_p spaces are defined analogously to the l_p spaces. For $p \ge 1$, the space $L_p[a, b]$ consists of those real-valued measurable functions x on the interval $[a, b]$ for which $|x(t)|^p$ is Lebesgue integrable. The norm on this space is defined as

$$\|x\|_p = \left(\int_a^b |x(t)|^p \, dt \right)^{1/p}.$$

Unfortunately, on this space $\|x\|_p = 0$ does not necessarily imply $x = \theta$ since x may be nonzero on a set of measure zero (such as a set consisting of a countable number of points). If, however, we do not distinguish

between functions that are equal almost everywhere,[2] then $L_p[a, b]$ is a normed linear space.

The space $R_p[a, b]$ is defined in an analogous manner but with attention restricted to functions x for which $|x(t)|$ is Riemann integrable. Since all functions encountered in applications in this book are Riemann integrable, we have little reason, at this point in the development, to prefer one of these spaces over the other. Readers unfamiliar with Lebesgue measure and integration theory lose no essential insight by considering the R_p spaces only. In the next section, however, when considering the notion of completeness, we find reason to prefer the L_p spaces over the R_p spaces.

The space $L_\infty[a, b]$ is the function space analog of l_∞. It is defined as the space of all Lebesgue measurable functions on $[a, b]$ which are bounded, except possibly on a set of measure zero. Again, in this space, two functions differing only on a set of measure zero are regarded as equivalent.

Roughly speaking, the norm of an element x in $L_\infty[a, b]$ is defined as $\sup_{a \le t \le b} |x(t)|$. This quantity is ambiguous, however, since an element x does not correspond uniquely to any given function but to a whole class of functions differing on a set of measure zero. The value $\sup_{a \le t \le b} |x(t)|$ is different for the different functions equivalent to x. The norm of a function in $L_\infty[a, b]$ is therefore defined as

$$\|x\|_\infty = \text{essential supremum of } |x(t)|$$
$$\equiv \operatorname*{infimum}_{y(t)=x(t) \text{ a.e.}} [\sup |y(t)|]$$

For brevity, we write $\|x\|_\infty = \operatorname{ess\,sup} |x(t)|$.

Example 1. Consider the function

$$x(t) = \begin{cases} 1 - t^2 & t \in [-1, 1], \quad t \neq 0 \\ 2 & t = 0 \end{cases}$$

shown in Figure 2.6. The supremum of this function is equal to 2 but the essential supremum is 1.

There are Hölder and Minkowski inequalities for the L_p spaces analogous to the corresponding results for the l_p spaces.

Theorem 3. (*The Hölder Inequality*) *If* $x \in L_p[a, b]$, $y \in L_q[a, b]$, $1/p + 1/q = 1$, $p, q > 1$, *then*

$$\int_a^b |x(t)y(t)|\, dt \le \|x\|_p \|y\|_q.$$

[2] Two functions are said to be equal almost everywhere (a.e.) on $[a, b]$ if they differ on a set of Lebesgue measure zero.

Equality holds if and only if

$$\left(\frac{|x(t)|}{\|x\|_p}\right)^p = \left(\frac{|y(t)|}{\|y\|_q}\right)^q$$

almost everywhere on $[a, b]$.

Theorem 4. (*The Minkowski Inequality*) *If* x *and* y *are in* $L_p[a, b]$, *then so is* $x + y$, *and* $\|x + y\|_p \leq \|x\|_p + \|y\|_p$.

The proofs of these inequalities are similar to those for the l_p spaces and are omitted here.

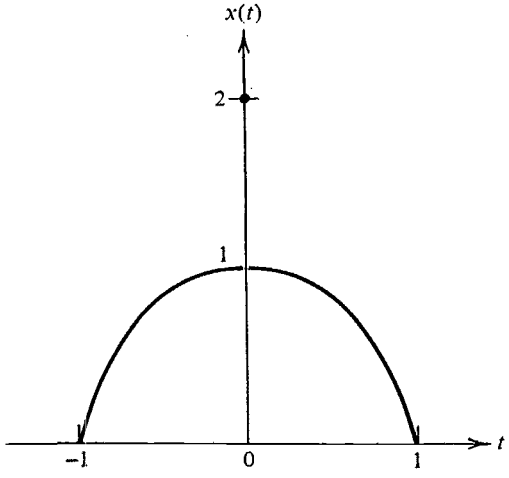

Figure 2.6 The function for Example 1

2.11 Banach Spaces

Definition. A sequence $\{x_n\}$ in a normed space is said to be a *Cauchy sequence* if $\|x_n - x_m\| \to 0$ as $n, m \to \infty$; i.e., given $\varepsilon > 0$, there is an integer N such that $\|x_n - x_m\| < \varepsilon$ for all $n, m > N$.

In a normed space, every convergent sequence is a Cauchy sequence since, if $x_n \to x$, then

$$\|x_n - x_m\| = \|x_n - x + x - x_m\| \leq \|x_n - x\| + \|x - x_m\| \to 0.$$

In general, however, a Cauchy sequence may not be convergent.

Normed spaces in which every Cauchy sequence is convergent are of particular interest in analysis; in such spaces it is possible to identify convergent sequences without explicitly identifying their limits. A space in which every Cauchy sequence has a limit (and is therefore convergent) is said to be complete.

Definition. A normed linear vector space X is *complete* if every Cauchy sequence from X has a limit in X. A complete normed linear vector space is called a *Banach space.*

We frequently take great care to formulate problems arising in applications as equivalent problems in Banach space rather than as problems in other, possibly incomplete, spaces. The principal advantage of Banach space in optimization problems is that when seeking an optimal vector maximizing a given objective, we often construct a sequence of vectors, each member of which is superior to the preceding members; the desired optimal vector is then the limit of the sequence. In order that the scheme be effective, there must be available a test for convergence which can be applied when the limit is unknown. The Cauchy criterion for convergence meets this requirement provided the underlying space is complete.

We now consider examples of incomplete normed spaces.

Example 1. Let X be the space of continuous functions on $[0, 1]$ with norm defined by $\|x\| = \int_0^1 |x(t)|\, dt$. One may readily verify that X is a normed linear space. Note that the space is *not* the space $C[0, 1]$ since the norm is different. We show that X is incomplete. Define a sequence of elements in X by the equation

$$x_n(t) = \begin{cases} 0 & \text{for } 0 \le t \le \frac{1}{2} - \frac{1}{n} \\ nt - \frac{n}{2} + 1 & \text{for } \frac{1}{2} - \frac{1}{n} \le t \le \frac{1}{2} \\ 1 & \text{for } t \ge \frac{1}{2} \end{cases}$$

This sequence of functions is illustrated in Figure 2.7. Each member of the sequence is a continuous function and thus a member of the space X. The sequence is Cauchy since, as is easily verified, $\|x_n - x_m\| = \frac{1}{2}|1/n - 1/m| \to 0$. It is obvious, however, that there is no *continuous* function to which the sequence converges.

Example 2. Let X be the vector space of finitely nonzero sequences $x = \{\xi_1, \xi_2, \ldots, \xi_n, 0, 0, \ldots\}$. Define the norm of an element x as $\|x\| = \max_i |\xi_i|$. Define a sequence of elements in X by

$$x_n = \left\{1, \frac{1}{2}, \frac{1}{3}, \cdots, \frac{1}{n-1}, 0, 0, \ldots\right\}.$$

Each x_n is in X and has its norm equal to unity. The sequence $\{x_n\}$ is a Cauchy sequence since, as is easily verified,

$$\|x_n - x_m\| = \max\{1/n, 1/m\} \to 0.$$

It is obvious, however, that there is no element of X (with a finite number of nonzero components) to which this sequence converges.

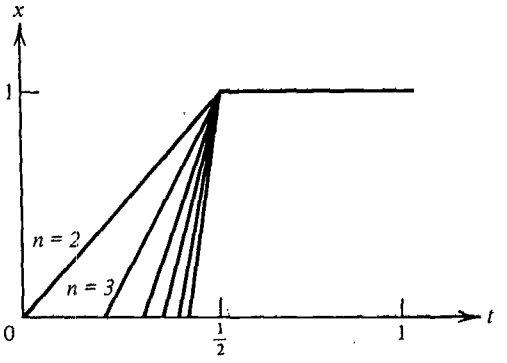

Figure 2.7 Sequence for Example 1

The space E^1, the real line, is the fundamental example of a complete space. It is assumed here that the completeness of E^1, a major topic in elementary analysis, is well known. The completeness of E^1 is used to establish the completeness of various other normed spaces.

We establish the completeness of several important spaces that are used throughout this book. For this purpose the following lemma is useful.

Lemma 1. *A Cauchy sequence is bounded.*

Proof. Let $\{x_n\}$ be a Cauchy sequence and let N be an integer such that $\|x_n - x_N\| < 1$ for $n > N$. For $n > N$, we have

$$\|x_n\| = \|x_n - x_N + x_N\| \leq \|x_N\| + \|x_n - x_N\| < \|x_N\| + 1. \quad \blacksquare$$

Example 3. **$C[0, 1]$ is a Banach space.** We have previously considered this space as a normed space. To prove that $C[0, 1]$ is complete, it is only necessary to show that every Cauchy sequence in $C[0, 1]$ has a limit. Suppose $\{x_n\}$ is a Cauchy sequence in $C[0, 1]$. For each fixed $t \in [0, 1]$, $|x_n(t) - x_m(t)| \leq \|x_n - x_m\| \to 0$, so $\{x_n(t)\}$ is a Cauchy sequence of real numbers. Since the set of real numbers is complete, there is a real number $x(t)$ to which the sequence converges; $x_n(t) \to x(t)$. Therefore, the functions x_n converge pointwise to the function x.

We prove next that this pointwise convergence is actually uniform in $t \in [0, 1]$, i.e., given $\varepsilon > 0$, there is an N such that $|x_n(t) - x(t)| < \varepsilon$ for all

$t \in [0, 1]$ and $n \geq N$. Given $\varepsilon > 0$, choose N such that $\|x_n - x_m\| < \varepsilon/2$ for $n, m > N$. Then for $n > N$

$$|x_n(t) - x(t)| \leq |x_n(t) - x_m(t)| + |x_m(t) - x(t)|$$
$$\leq \|x_n - x_m\| + |x_m(t) - x(t)|.$$

By choosing m sufficiently large (which may depend on t), each term on the right can be made smaller than $\varepsilon/2$ so $|x_n(t) - x(t)| < \varepsilon$ for $n > N$.

We must still prove that the function x is continuous and that the sequence $\{x_n\}$ converges to x in the norm of $C[0, 1]$. To prove the continuity of x, fix $\varepsilon > 0$. For every δ, t, and n,

$$|x(t + \delta) - x(t)| \leq |x(t + \delta) - x_n(t + \delta)|$$
$$+ |x_n(t + \delta) - x_n(t)| + |x_n(t) - x(t)|.$$

Since $\{x_n\}$ converges uniformly to x, n may be chosen to make both the first and the last terms less than $\varepsilon/3$ for all δ. Since x_n is continuous, δ may be chosen to make the second term less than $\varepsilon/3$. Therefore, x is continuous. The convergence of x_n to x in the $C[0, 1]$ norm is a direct consequence of the uniform convergence.

It is instructive to reconcile the completeness of $C[0, 1]$ with Example 1 in which a sequence of continuous functions was shown to be Cauchy but nonconvergent with respect to the integral norm. The difference is that, with respect to the $C[0, 1]$ norm, the sequence defined in Example 1 is not Cauchy. The reader may find it useful to compare these two cases in detail.

***Example 4.* l_p, $1 \leq p \leq \infty$ is a Banach space.** Assume first that $1 \leq p < \infty$. Let $\{x_n\}$ be a Cauchy sequence in l_p. Then, if $x_n = \{\xi_1^n, \xi_2^n, \ldots\}$, we have

$$|\xi_k^n - \xi_k^m| \leq \left\{ \sum_{i=1}^{\infty} |\xi_i^n - \xi_i^m|^p \right\}^{1/p} \to 0$$

Hence, for each k, the sequence $\{\xi_k^n\}$ is a Cauchy sequence of real numbers and, therefore, converges to a limit ξ_k. We show now that the element $x = \{\xi_1, \xi_2, \ldots\}$ is a vector in l_p. According to Lemma 1, there is a constant M which bounds the sequence $\{x_n\}$. Hence, for all n and k

$$\sum_{i=1}^{k} |\xi_i^n|^p \leq \|x_n\|^p \leq M^p.$$

Since the left member of the inequality is a finite sum, the inequality remains valid as $n \to \infty$; therefore,

$$\sum_{i=1}^{k} |\xi_i|^p \leq M^p.$$

Since this inequality holds uniformly in k, it follows that

$$\sum_{i=1}^{\infty} |\xi_i|^p \leq M^p$$

and $x \in l_p$; the norm of x is bounded by $\|x\| \leq M$.

It remains to be shown that the sequence $\{x_n\}$ converges to x in the l_p norm. Given $\varepsilon > 0$, there is an N such that

$$\sum_{i=1}^{k} |\xi_i^n - \xi_i^m|^p \leq \|x_n - x_m\|^p \leq \varepsilon$$

for $n, m > N$ and for each k. Letting $m \to \infty$ in the finite sum on the left, we conclude that

$$\sum_{i=1}^{k} |\xi_i^n - \xi_i|^p \leq \varepsilon$$

for $n > N$. Now letting $k \to \infty$, we deduce that $\|x_n - x\|^p \leq \varepsilon$ for $n > N$ and, therefore, $x_n \to x$. This completes the proof for $1 \leq p < \infty$.

Now let $\{x_n\}$ be a Cauchy sequence in l_∞. Then $|\xi_k^n - \xi_k^m| \leq \|x_n - x_m\| \to 0$. Hence, for each k there is a real number ξ_k such that $\xi_k^n \to \xi_k$. Furthermore, this convergence is uniform in k. Let $x = \{\xi_1, \xi_2, \ldots\}$. Since $\{x_n\}$ is Cauchy, there is a constant M such that for all n, $\|x_n\| \leq M$. Therefore, for each k and each n, we have $|\xi_k^n| \leq \|x_n\| \leq M$ from which it follows that $x \in l_\infty$ and $\|x\| \leq M$.

The convergence of x_n to x follows directly from the uniform convergence of $\xi_k^n \to \xi_k$.

Example 5. **$L_p[0, 1]$, $1 \leq p \leq \infty$ is a Banach space.** We do not prove the completeness of the L_p spaces because the proof requires a fairly thorough familiarity with Lebesgue integration theory. Consider instead the space R_p consisting of all functions x on $[0, 1]$ for which $|x|^p$ is Riemann integrable with norm defined as

$$\|x\| = \left\{\int_0^1 |x(t)|^p \, dt\right\}^{1/p}.$$

The normed space R_p is incomplete. It may be completed by adjoining to it certain additional functions derived from Cauchy sequences in R_p. In this way, R_p is imbedded in a larger normed space which is complete. The smallest complete space containing R_p is L_p. A general method for completing a normed space is discussed in Problem 15.

Example 6. Given two normed spaces X, Y, we consider the product space $X \times Y$ consisting of ordered pairs (x, y) as defined in Section 2.2. The space $X \times Y$ can be normed in several ways such as $\|(x, y)\| = \|x\| + \|y\|$

or $\|(x, y)\| = \max \{\|x\|, \|y\|\}$ but, unless specifically noted otherwise, we define the product norm as $\|(x, y)\| = \|x\| + \|y\|$. It is simple to show that if X and Y are Banach spaces, the product space $X \times Y$ with the product norm is also a Banach space.

2.12 Complete Subsets

The definition of completeness has an obvious extension to subsets of a normed space; a subset is complete if every Cauchy sequence from the subset converges to a limit in the subset. The following theorem states that completeness and closure are equivalent in a Banach space. This is not so in general normed space since, for example, a normed space is always closed but not necessarily complete.

Theorem 1. *In a Banach space a subset is complete if and only if it is closed.*

Proof. A complete subset is obviously closed since every convergent (and hence Cauchy) sequence has a limit in the subset. A Cauchy sequence from a closed subset has a limit somewhere in the Banach space. By closure the limit must be in the subset. ▮

The following theorem is of great importance in many applications.

Theorem 2. *In a normed linear space, any finite-dimensional subspace is complete.*

Proof. The proof is by induction on the dimension of the subspace. A one-dimensional subspace is complete since, in such a subspace, all elements have the form $x = \alpha e$ where α is an arbitrary scalar and e is a fixed vector. Convergence of a sequence $\alpha_n e$ is equivalent to convergence of the sequence of scalars $\{\alpha_n\}$ and, hence, completeness follows from the completeness of E.

Assume that the theorem is true for subspaces of dimension $N - 1$. Let X be a normed space and M an N-dimensional subspace of X. We show that M is complete.

Let $\{e_1, e_2, \ldots, e_N\}$ be a basis for M. For each k, define

$$\delta_k = \inf_{\alpha_j's} \|e_k - \sum_{j \neq k} \alpha_j e_j\|.$$

The number δ_k is the distance from the vector e_k to the subspace M_k generated by the remaining $N - 1$ basis vectors. The number δ_k is greater than zero because otherwise a sequence of vectors in the $N - 1$ dimensional subspace M_k could be constructed converging to e_k. Such a sequence cannot exist since M_k is complete by the induction hypothesis.

Define $\delta > 0$ as the minimum of the δ_k, $k = 1, 2, \ldots, N$. Suppose that $\{x_n\}$ is a Cauchy sequence in M. Each x_n has a unique representation as

$$x_n = \sum_{i=1}^{N} \lambda_i^n e_i.$$

For arbitrary n, m

$$\|x_n - x_m\| = \left\| \sum_{i=1}^{N} (\lambda_i^n - \lambda_i^m) e_i \right\| \geq |\lambda_k^n - \lambda_k^m| \delta$$

for each k, $1 \leq k \leq N$. Since $\|x_n - x_m\| \to 0$, each $|\lambda_k^n - \lambda_k^m| \to 0$. Thus, $\{\lambda_k^n\}_{n=1}^{\infty}$ is a Cauchy sequence of scalars and hence convergent to a scalar λ_k. Let $x = \sum_{k=1}^{n} \lambda_k e_k$. Obviously, $x \in M$. We show that $x_n \to x$. For all n, we have

$$\|x_n - x\| = \left\| \sum_{k=1}^{N} (\lambda_k^n - \lambda_k) e_k \right\| \leq N \cdot \max_{1 \leq k \leq N} |\lambda_k^n - \lambda_k| \cdot \|e_k\|,$$

but since $|\lambda_k^n - \lambda_k| \to 0$ for all k, $\|x_n - x\| \to 0$. Thus, $\{x_n\}$ converges to $x \in M$. ∎

*2.13 Extreme Values of Functionals and Compactness

Optimization theory is largely concerned with the maximization or minimization of real functionals over a given subset; indeed, a major portion of this book is concerned with principles for finding the points at which a given functional attains its maximum. A more fundamental question, however, is whether a functional has a maximum on a given set. In many cases the answer to this is easily established by inspection, but in others it is by no means obvious.

In finite-dimensional spaces the well-known Weierstrass theorem, which states that a continuous function defined on a closed and bounded (compact) set has a maximum and a minimum, is of great utility. Usually, this theorem alone is sufficient to establish the existence of a solution to a given optimization problem.

In this section we generalize the Weierstrass theorem to compact sets in a normed space, thereby obtaining a simple and yet general result applicable to infinite-dimensional problems. Unfortunately, however, the restriction to compact sets is so severe in infinite-dimensional normed spaces that the Weierstrass theorem can in fact only be employed in the minority of optimization problems. The theorem, however, deserves special attention in optimization theory if only because of the finite-dimensional version. The interested reader should also consult Section 5.10.

Actually, to prove the Weierstrass theorem it is not necessary to assume continuity of the functional but only upper semicontinuity. This added generality is often of great utility.

Definition. A (real-valued) functional f defined on a normed space X is said to be *upper semicontinuous* at x_0 if, given $\varepsilon > 0$, there is a $\delta > 0$ such that $f(x) - f(x_0) < \varepsilon$ for $\|x - x_0\| < \delta$. A functional f is said to be *lower semicontinuous* at x_0 if $-f$ is upper semicontinuous at x_0.

As the reader may verify, an equivalent definition is that f is upper semicontinuous at x_0 if[3] $\limsup_{x \to x_0} f(x) \leq f(x_0)$. Clearly, if f is both upper and lower semicontinuous, it is continuous.

Definition. A set K in a normed space X is said to be *compact* if, given an arbitrary sequence $\{x_i\}$ in K, there is a subsequence $\{x_{i_n}\}$ converging to an element $x \in K$.

In finite dimensions, compactness is equivalent to being closed and bounded, but, as is shown below, this is not true in general normed space. Note, however, that a compact set K must be complete since any Cauchy sequence from K must have a limit in K.

Theorem 1. (*Weierstrass*) *An upper semicontinuous functional on a compact subset K of a normed linear space X achieves a maximum on K.*

Proof. Let $M = \sup_{x \in K} f(x)$ (we allow the possibility $M = \infty$). There is a sequence $\{x_i\}$ from K such that $f(x_i) \to M$. Since K is compact, there is a convergent subsequence $x_{i_n} \to x \in K$. Clearly, $f(x_{i_n}) \to M$ and, since f is upper semicontinuous, $f(x) \geq \lim f(x_{i_n}) = M$. Thus, since $f(x)$ must be finite, we conclude that $M < \infty$ and that $f(x) = M$. ∎

We offer now an example of a continuous functional on the unit sphere of $C[0, 1]$ which does not attain a maximum (thus proving that the unit sphere is not compact).

Example 1. Let the functional f be defined on $C[0, 1]$ by

$$f(x) = \int_0^{1/2} x(t)\, dt - \int_{1/2}^1 x(t)\, dt.$$

It is easily verified that f is continuous since, in fact, $|f(x)| \leq \|x\|$. The supremum of f over the unit sphere in $C[0, 1]$ is 1, but no continuous function of norm less than unity achieves this supremum. (If the problem

[3] The lim sup of a functional on a normed space is the obvious extension of the corresponding definition for functions of a real variable.

were formulated in $L_\infty[0, 1]$, the supremum would be achieved by a function discontinuous at $t = \frac{1}{2}$.)

Example 2. Suppose that in the above example we restrict our attention to those continuous functions in $C[0, 1]$ within the unit sphere which are polynomials of degree n or less. The set of permissible elements in $C[0, 1]$ is now a closed, bounded subset of the finite-dimensional space of n-th degree polynomials and is therefore compact. Thus Weierstrass's theorem guarantees the existence of a maximizing vector from this set.

*2.14 Quotient Spaces

Suppose we select a subspace M from a vector space X and generate linear varieties V in X by translations of M. The linear varieties obtained can be regarded as the elements of a new vector space called the quotient space of X modulo M and denoted X/M. If, for example M is a plane through the origin in three-dimensional space, X/M consists of the family of planes parallel to M. We formalize this definition below.

Definition. Let M be a subspace of a vector space X. Two elements $x_1, x_2 \in X$ are said to be *equivalent modulo* M if $x_1 - x_2 \in M$. In this case, we write $x_1 \equiv x_2$.

This equivalence relation partitions the space X into disjoint subsets, or classes, of equivalent elements: namely, the linear varieties that are distinct translates of the subspace M. These classes are often called the *cosets* of M. Given an arbitrary element $x \in X$, it belongs to a unique coset of M which we denote[4] by $[x]$.

Definition. Let M be a subspace of a vector space X. The *quotient space* X/M consists of all cosets of M with addition and scalar multiplication defined by $[x_1] + [x_2] = [x_1 + x_2]$, $\alpha[x] = [\alpha x]$.

Several things, which we leave to the reader, need to be verified in order to justify the above definition. It must be shown that the definitions for addition and scalar multiplication are independent of the choice of representative elements, and the axioms for a vector space must be verified. These matters are easily proved, however, and it is not difficult to see that addition of cosets $[x_1]$ and $[x_2]$ merely amounts to addition of the corresponding linear varieties regarded as sets in X. Likewise, multiplication of a coset by a scalar α (except for $\alpha = 0$) amounts to multiplication of the corresponding linear variety by α.

[4] The notation $[x]$ is also used for the subspace generated by x but the usage is always clear from context.

Suppose now that X is a normed space and that M is a closed subspace of X. We define the norm of a coset $[x] \in X/M$ by

$$\|[x]\| = \inf_{m \in M} \|x + m\|,$$

i.e., $\|[x]\|$ is the infimum of the norms of all elements in the coset $[x]$. The assumption that M is closed insures that $\|[x]\| > 0$ if $[x] \neq \theta$. Satisfaction of the other two axioms for a norm is easily verified. In the case of X being two dimensional, M one dimensional, and X/M consisting of parallel lines, the quotient norm of one of the lines is the minimum distance of the line from the origin.

Proposition 1. *Let X be a Banach space, M a closed subspace of X, and X/M the quotient space with the quotient norm defined as above. Then X/M is also a Banach space.*

The proof is left to the reader.

*2.15 Denseness and Separability

We conclude this chapter by introducing one additional topological concept, that of denseness.

Definition. A set D is said to be *dense* in a normed space X if for each element $x \in X$ and each $\varepsilon > 0$ there exists $d \in D$ with $\|x - d\| < \varepsilon$.

If D is dense in X, there are points of D arbitrarily close to each $x \in X$. Thus, given x, a sequence can be constructed from D which converges to x. It follows that equivalent to the above definition is the statement: D is dense in X if $\bar{D}$, the closure of D, is X.

The definition converse to the above is that of a nowhere dense set.

Definition. A set E is said to be *nowhere dense* in a normed space X if $\bar{E}$ contains no open set.

The classic example of a dense set is the set of rationals in the real line. Another example, provided by the well-known Weierstrass approximation theorem, is that the space of polynomials is dense in the space $C[a, b]$.

Definition. A normed space is *separable* if it contains a countable dense set.

Most, but not all, of the spaces considered in this book are separable

Example 1. The space E^n is separable. The collection of vectors $x = (\xi_1, \xi_2, \ldots, \xi_n)$ having rational components is countable and dense in E^n.

Example 2. The l_p spaces, $1 \leq p < \infty$, are separable. To prove separability, let D be the set consisting of all finitely nonzero sequences with rational components. D is easily seen to be countable. Let $x = \{\xi_1, \xi_2, \ldots\}$ exist in l_p, $1 \leq p < \infty$, and fix $\varepsilon > 0$. Since $\sum_{i=1}^{\infty} |\xi_i|^p < \infty$, there is an N such that $\sum_{i=N+1}^{\infty} |\xi_i|^p < \varepsilon/2$. For each k, $1 \leq k \leq N$, let r_k be a rational such that $|\xi_k - r_k|^p < \varepsilon/2N$; let $d = \{r_1, r_2, \ldots, r_N, 0, 0, \ldots\}$. Clearly, $d \in D$ and $\|x - d\|^p < \varepsilon$. Thus, D is dense in l_p.

The space l_∞ is not separable.

Example 3. The space $C[a, b]$ is separable. Indeed, the countable set consisting of all polynomials with rational coefficients is dense in $C[a, b]$. Given $x \in C[a, b]$ and $\varepsilon > 0$, it follows from the well-known Weierstrass approximation theorem that a polynomial p can be found such that $|x(t) - p(t)| < \varepsilon/2$ for all $t \in [a, b]$. Clearly, there is another polynomial r with rational coefficients such that $|p(t) - r(t)| < \varepsilon/2$ for all $t \in [a, b]$ can be constructed by changing each coefficient by less than $\varepsilon/2N$ where $N - 1$ is the order of the polynomial p). Thus

$$\|x - r\| = \max_{t \in [a, b]} |x(t) - r(t)| \leq \max |x(t) - p(t)| + \max |p(t) - r(t)|$$

$$\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Example 4. The L_p spaces $1 \leq p < \infty$ are separable but L_∞ is not separable. The particular space L_2 is considered in great detail in Chapter 3.

2.16 Problems

1. Prove Proposition 1, Section 2.2.
2. Show that in a vector space $(-\alpha)x = \alpha(-x) = -(\alpha x)$, $\alpha(x - y) = \alpha x - \alpha y$, $(\alpha - \beta)x = \alpha x - \beta x$.
3. Let M and N be subspaces in a vector space. Show that $[M \cup N] = M + N$.
4. A *convex combination* of the vectors $x_1, x_2, \ldots, x_n$ is a linear combination of the form $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$ where $\alpha_i \geq 0$, for each i; and $\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$. Given a set S in a vector space, let K be the set of vectors consisting of all convex combinations from S. Show that $K = \text{co}(S)$.
5. Let C and D be convex cones in a vector space. Show that $C \cap D$ and $C + D$ are convex cones.
6. Prove that the union of an arbitrary collection of open sets is open and that the intersection of a finite collection of open sets is open.

7. Prove that the intersection of an arbitrary collection of closed sets is closed and that the union of a finite collection of closed sets is closed.
8. Show that the closure of a set S in a normed space is the smallest closed set containing S.
9. Let X be a normed linear space and let $x_1, x_2, \ldots, x_n$ be linearly independent vectors from X. For fixed $y \in X$, show that there are coefficients $a_1, a_2, \ldots, a_n$ minimizing $\|y - a_1x_1 - a_2x_2 \cdots - a_nx_n\|$.
10. A normed space is said to be *strictly normed* if $\|x + y\| = \|x\| + \|y\|$ implies that $y = \theta$ or $x = \alpha y$ for some α.
    (a) Show that $L_p[0, 1]$ is strictly normed for $1 < p < \infty$.
    (b) Show that if X is strictly normed the solution to Problem 9 is unique.
11. Prove the Hölder inequality for $p = 1, \infty$.
12. Suppose $x = \{\xi_1, \xi_2, \ldots, \xi_n, \ldots\} \in l_{p_0}$ for some p_0, $1 \leq p_0 < \infty$.
    (a) Show that $x \in l_p$ for all $p \geq p_0$.
    (b) Show that $\|x\|_\infty = \lim_{p \to \infty} \|x\|_p$.
13. Prove that the normed space $D[a, b]$ is complete.
14. Two vector spaces, X and Y, are said to be *isomorphic* if there is a one-to-one mapping T of X onto Y such that $T(\alpha_1x_1 + \alpha_2x_2) = \alpha_1T(x_1) + \alpha_2T(x_2)$. Show that any real n-dimensional space is isomorphic to E^n.
15. Two normed linear spaces, X and Y, are said to be *isometrically isomorphic* if they are isomorphic and if the corresponding one-to-one mapping T satisfies $\|T(x)\| = \|x\|$. The object of this problem is to show that any normed space X is isometrically isomorphic to a dense subset of a Banach space $\hat{X}$.
    (a) Let $\overline{X}$ be the set of all Cauchy sequences $\{x_n\}$ from X with addition and scalar multiplication defined coordinatewise. Show that $\overline{X}$ is a linear space.
    (b) If $y = \{x_n\} \in \overline{X}$, define $\|y\| = \sup_n \|x_n\|$. Show that with this definition, $\overline{X}$ becomes a normed space.
    (c) Let M be the subspace of $\overline{X}$ consisting of all Cauchy sequences convergent to zero, and let $\hat{X} = \overline{X}/M$. Show that $\hat{X}$ has the required properties.
16. Let S, T be open sets in the normed spaces X and Y, respectively. Show that $S \times T = \{(s, t) : s \in S, t \in T\}$ is an open subset of $X \times Y$ under the usual product norm.
17. Let M be an m-dimensional subspace of an n-dimensional vector space X. Show that X/M is $(n - m)$ dimensional.
18. Let X be n dimensional and Y be m dimensional. What is the dimension of $X \times Y$?

19. A real-valued functional $|x|$, defined on a vector space X, is called a *seminorm* if:
    1. $|x| \geq 0$ all $x \in X$
    2. $|\alpha x| = |\alpha| \cdot |x|$
    3. $|x + y| \leq |x| + |y|$.

    Let $M = \{x : |x| = 0\}$ and show that the space X/M with $\|[x]\| = \inf_{m \in M} |x + m|$ is normed.
20. Let X be the space of all functions x on $[0, 1]$ which vanish at all but a countable number of points and for which

$$\|x\| = \sum_{n=1}^{\infty} |x(t_n)| < \infty,$$

    where the t_n are the points at which x does not vanish.
    (a) Show that X is a Banach space.
    (b) Show that X is *not* separable.
21. Show that the Banach space l_∞ is not separable.

REFERENCES

There are a number of excellent texts on linear spaces that can be used to supplement this chapter.

§2.2. For the basic concepts of vector space, an established standard is Halmos [68].

§2.4. Convexity is an important aspect of a number of mathematical areas. For an introductory discussion, consult Eggleston [47] or, more advanced, Valentine [148].

§2.6–9. For general discussions of functional analysis, refer to any of the following. They are listed in approximate order of increasing level: Kolmogorov and Fomin [88], Simmons [139], Luisternik and Sobolev [101], Goffman and Pedrick [59], Kantorovich and Akilov [79], Taylor [145], Yosida [157], Riesz and Sz-Nagy [123], Edwards [46], Dunford and Schwartz [45], and Hille and Phillips [73].

§2.10. For a readable discussion of l_p and L_p spaces and a general reference on analysis and integration theory, see Royden [131].

§2.13. For the Weierstrass approximation theorem, see Apostol [10].

3

HILBERT SPACE

3.1 Introduction

Every student of high school geometry learns that the shortest distance from a point to a line is given by the perpendicular from the point to the line. This highly intuitive result is easily generalized to the problem of finding the shortest distance from a point to a plane; furthermore one might reasonably conjecture that in n-dimensional Euclidean space the shortest vector from a point to a subspace is orthogonal to the subspace. This is, in fact, a special case of one of the most powerful and important optimization principles—the projection theorem.

The key concept in this observation is that of orthogonality; a concept which is not generally available in normed space but which is available in Hilbert space. A Hilbert space is simply a special form of normed space having an inner product defined which is analogous to the dot product of two vectors in analytic geometry. Two vectors are then defined as orthogonal if their inner product is zero.

Hilbert spaces, equipped with their inner products, possess a wealth of structural properties generalizing many of our geometrical insights for two and three dimensions. Correspondingly, these structural properties imply a wealth of analytical results applicable to problems formulated in Hilbert space. The concepts of orthonormal bases, Fourier series, and least-squares minimization all have natural settings in Hilbert space.

PRE-HILBERT SPACES

3.2 Inner Products

Definition. A *pre-Hilbert space* is a linear vector space X together with an *inner product* defined on $X \times X$. Corresponding to each pair of vectors x, y in X the inner product $(x \mid y)$ of x and y is a scalar. The inner product satisfies the following axioms:

1. $(x \mid y) = \overline{(y \mid x)}$.
2. $(x + y \mid z) = (x \mid z) + (y \mid z)$.

3. $(\lambda x \mid y) = \lambda(x \mid y)$.
4. $(x \mid x) \geq 0$ and $(x \mid x) = 0$ if and only if $x = \theta$.

The bar on the right side of axiom 1 denotes complex conjugation. The axiom itself guarantees that $(x \mid x)$ is real for each x. Together axioms 2 and 3 imply that the inner product is linear in the first entry. We distinguish between real and complex pre-Hilbert spaces according to whether the underlying vector space is real or complex. In the case of real pre-Hilbert spaces, it is required that the inner product be real valued; it then follows that the inner product is linear in both entries. In this book the pre-Hilbert spaces are almost exclusively considered to be real.

The quantity $\sqrt{(x \mid x)}$ is denoted $\|x\|$, and our first objective is to verify that it is indeed a norm in the sense of Chapter 2. Axioms 1 and 3 together give $\|\alpha x\| = |\alpha| \, \|x\|$ and axiom 4 gives $\|x\| > 0$, $x \neq \theta$. It is shown in Proposition 1 that $\| \ \|$ satisfies the triangle inequality and, hence, defines a norm on the pre-Hilbert space.

Before proving the triangle inequality, it is first necessary to prove an important lemma which is fundamental throughout this chapter.

Lemma 1. *(The Cauchy-Schwarz Inequality) For all x, y in an inner product space, $|(x \mid y)| \leq \|x\| \|y\|$. Equality holds if and only if $x = \lambda y$ or $y = \theta$.*

Proof. If $y = \theta$, the inequality holds trivially. Therefore, assume $y \neq \theta$. For all scalars λ, we have

$$0 \leq (x - \lambda y \mid x - \lambda y) = (x \mid x) - \lambda(y \mid x) - \bar{\lambda}(x \mid y) + |\lambda|^2 (y \mid y).$$

In particular, for $\lambda = (x \mid y)/(y \mid y)$, we have

$$0 \leq (x \mid x) - \frac{|(x \mid y)|^2}{(y \mid y)},$$

or

$$|(x \mid y)| \leq \sqrt{(x \mid x)(y \mid y)} = \|x\| \|y\|. \ \blacksquare$$

Proposition 1. *On a pre-Hilbert space X the function $\|x\| = \sqrt{(x \mid x)}$ is a norm.*

Proof. The only requirement for a norm which has not already been established is the triangle inequality. For any $x, y \in X$, we have

$$\begin{aligned} \|x + y\|^2 &= (x + y \mid x + y) = (x \mid x) + (x \mid y) + (y \mid x) + (y \mid y) \\ &\leq \|x\|^2 + 2|(x \mid y)| + \|y\|^2. \end{aligned}$$

By the Cauchy-Schwarz inequality, this becomes

$$\|x + y\|^2 \leq \|x\|^2 + 2\|x\| \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.$$

The square root of the above inequality is the desired result. ∎

Several of the normed spaces that we have previously considered can be converted to pre-Hilbert spaces by introducing an appropriate inner product.

Example 1. The space E^n consisting of n-tuples of real numbers is a pre-Hilbert space with the inner product of the vector $x = (\xi_1, \xi_2, \ldots, \xi_n)$ and the vector $y = (\eta_1, \eta_2, \ldots, \eta_n)$ defined as $(x \mid y) = \sum_{i=1}^{n} \xi_i \eta_i$. In this case it is clear that $(x \mid y) = (y \mid x)$ and that $(x \mid y)$ is linear in both entries. The norm defined as $\sqrt{(x \mid x)}$ is

$$\|x\| = \left(\sum_{i=1}^{n} |\xi_i|^2 \right)^{1/2}$$

which is the Euclidean norm for E^n.

Example 2. The (real) space l_2 becomes a pre-Hilbert space with the inner product of the vectors $x = \{\xi_1, \xi_2, \ldots\}$ and $y = \{\eta_1, \eta_2, \ldots\}$ defined as $(x \mid y) = \sum_{i=1}^{\infty} \xi_i \eta_i$. The Hölder inequality for l_2 (which becomes the Cauchy-Schwarz inequality) guarantees that $|(x \mid y)| \leq \|x\| \cdot \|y\|$ and, thus, the inner product has finite value. The norm defined by $\sqrt{(x \mid x)}$ is the usual l_2 norm.

Example 3. The (real) space $L_2[a, b]$ is a pre-Hilbert space with the inner product defined as $(x \mid y) = \int_a^b x(t)y(t)\, dt$. Again the Hölder inequality guarantees that $(x \mid y)$ is finite.

Example 4. The space of polynomial functions on $[a, b]$ with inner product $(x \mid y) = \int_a^b x(t)y(t)\, dt$ is a pre-Hilbert space. Obviously, this space is a subspace of the pre-Hilbert space $L_2[a, b]$.

There are various properties of inner products which are a direct consequence of the definition. Some of these are useful in later developments.

Lemma 2. *In a pre-Hilbert space the statement* $(x \mid y) = 0$ *for all* y *implies that* $x = \theta$.

Proof. Putting $y = x$ implies $(x \mid x) = 0$. ∎

Lemma 3. (*The Parallelogram Law*) *In a pre-Hilbert space*

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

Proof. The proof is made by direct expansion of the norms in terms of the inner product. ∎

This last result is a generalization of a result for parallelograms in two-dimensional geometry. The sum of the squares of the lengths of the diagonals of a parallelogram is equal to twice the sum of the squares of two adjacent sides. See Figure 3.1.

Since a pre-Hilbert space is a special kind of normed linear space, the concepts of convergence, closure, completeness, etc., apply in these spaces.

Definition. A complete pre-Hilbert space is called a *Hilbert space.*

A Hilbert space, then, is a Banach space equipped with an inner product which induces the norm. The spaces E^n, l_2, and $L_2[a, b]$ are all Hilbert spaces. Inner products enjoy the following useful continuity property.

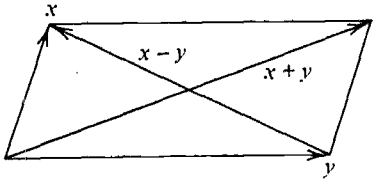

Figure 3.1 The parallelogram law

Lemma 4. *(Continuity of the Inner Product) Suppose that* $x_n \to x$ *and* $y_n \to y$ *in a pre-Hilbert space. Then* $(x_n \mid y_n) \to (x \mid y)$.

Proof. Since the sequence $\{x_n\}$ is convergent, it is bounded; say $\|x_n\| \leq M$. Now

$$\begin{aligned}|(x_n \mid y_n) - (x \mid y)| &= |(x_n \mid y_n) - (x_n \mid y) \\ &\quad + (x_n \mid y) - (x \mid y)| \leq |(x_n \mid y_n - y)| + |(x_n - x \mid y)|.\end{aligned}$$

Applying the Cauchy-Schwarz inequality, we obtain

$$|(x_n \mid y_n) - (x \mid y)| \leq \|x_n\| \, \|y_n - y\| + \|x_n - x\| \, \|y\|.$$

Since $\|x_n\|$ is bounded,

$$|(x_n \mid y_n) - (x \mid y)| \leq M\|y_n - y\| + \|x_n - x\| \, \|y\| \to 0. \quad \blacksquare$$

3.3 The Projection Theorem

We get a lot of analytical mileage from the following definition.

Definition. In a pre-Hilbert space two vectors x, y are said to be *orthogonal* if $(x \mid y) = 0$. We symbolize this by $x \perp y$. A vector x is said to be orthogonal to a set S (written $x \perp S$) if $x \perp s$ for each $s \in S$.

The concept of orthogonality has many of the consequences in pre-Hilbert spaces that it has in plane geometry. For example, the Pythagorean theorem is true in pre-Hilbert spaces.

Lemma 1. *If* $x \perp y$, *then* $\|x + y\|^2 = \|x\|^2 + \|y\|^2$.

Proof.

$$\|x+y\|^2 = (x+y\,|\,x+y) = \|x\|^2 + (x\,|\,y) + (y\,|\,x) + \|y\|^2 = \|x\|^2 + \|y\|^2. \blacksquare$$

We turn now to our first optimization problem and the projection theorem which characterizes its solution. We prove two slightly different versions of the theorem: one valid in an arbitrary pre-Hilbert space and another, with a stronger conclusion, valid in Hilbert space.

The optimization problem considered is this: Given a vector x in a pre-Hilbert space X and a subspace M in X, find the vector $m \in M$ closest to x in the sense that it minimizes $\|x-m\|$. Of course, if x itself lies in M, the solution is trivial. In general, however, three important questions must be answered for a complete solution to the problem. First, is there a vector $m \in M$ which minimizes $\|x-m\|$, or is there no m that is at least as good as all others? Second, is the solution unique? And third, what is the solution or how is it characterized? We answer these questions now.

Theorem 1. *Let X be a pre-Hilbert space, M a subspace of X, and x an arbitrary vector in X. If there is a vector $m_0 \in M$ such that $\|x-m_0\| \leq \|x-m\|$ for all $m \in M$, then m_0 is unique. A necessary and sufficient condition that $m_0 \in M$ be a unique minimizing vector in M is that the error vector $x-m_0$ be orthogonal to M.*

Proof. We show first that if m_0 is a minimizing vector, then $x-m_0$ is orthogonal to M. Suppose to the contrary that there is an $m \in M$ which is not orthogonal to $x-m_0$. Without loss of generality, we may assume that $\|m\| = 1$ and that $(x-m_0\,|\,m) = \delta \neq 0$. Define the vector m_1 in M as $m_1 = m_0 + \delta m$. Then

$$\begin{aligned}\|x-m_1\|^2 = \|x-m_0-\delta m\|^2 &= \|x-m_0\|^2 - (x-m_0\,|\,\delta m) \\ &\quad -(\delta m\,|\,x-m_0) + |\delta|^2 \\ &= \|x-m_0\|^2 - |\delta|^2 < \|x-m_0\|^2.\end{aligned}$$

Thus, if $x-m_0$ is not orthogonal to M, m_0 is not a minimizing vector.

We show now that if $x-m_0$ is orthogonal to M, then m_0 is a unique minimizing vector. For any $m \in M$, the Pythagorean theorem gives

$$\|x-m\|^2 = \|x-m_0+m_0-m\|^2 = \|x-m_0\|^2 + \|m_0-m\|^2.$$

Thus, $\|x-m\| > \|x-m_0\|$ for $m \neq m_0$. $\blacksquare$

The three-dimensional version of this theorem is illustrated in Figure 3.2.

We still have not established the existence of the minimizing vector. We have shown that if it exists, it is unique and that $x-m_0$ is orthogonal to the subspace M. By slightly strengthening the hypotheses, we can also guarantee the existence of the minimizing vector.

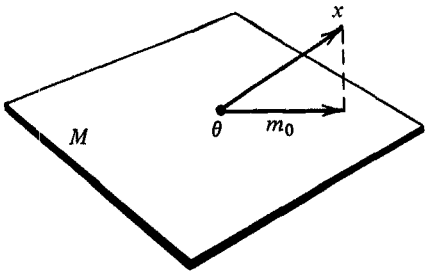

Figure 3.2 The projection theorem

Theorem 2. (*The Classical Projection Theorem*) *Let H be a Hilbert space and M a closed subspace of H. Corresponding to any vector $x \in H$, there is a unique vector $m_0 \in M$ such that $\|x - m_0\| \leq \|x - m\|$ for all $m \in M$. Furthermore, a necessary and sufficient condition that $m_0 \in M$ be the unique minimizing vector is that $x - m_0$ be orthogonal to M.*

Proof. The uniqueness and orthogonality have been established in Theorem 1. It is only required to establish the existence of the minimizing vector.

If $x \in M$, then $m_0 = x$ and everything is settled. Let us assume $x \notin M$ and define $\delta = \inf_{m \in M} \|x - m\|$. We wish to produce an $m_0 \in M$ with $\|x - m_0\| = \delta$. For this purpose, let $\{m_i\}$ be a sequence of vectors in M such that $\|x - m_i\| \to \delta$. Now, by the parallelogram law,

$$\|(m_j - x) + (x - m_i)\|^2 + \|(m_j - x) - (x - m_i)\|^2 = 2\|m_j - x\|^2 + 2\|x - m_i\|^2.$$

Rearranging, we obtain

$$\|m_j - m_i\|^2 = 2\|m_j - x\|^2 + 2\|x - m_i\|^2 - 4\left\|x - \frac{m_i + m_j}{2}\right\|^2.$$

For all i, j the vector $(m_i + m_j)/2$ is in M since M is a linear subspace. Therefore, by definition of δ, $\|x - (m_i + m_j)/2\| \geq \delta$ and we obtain

$$\|m_j - m_i\|^2 \leq 2\|m_j - x\|^2 + 2\|x - m_i\|^2 - 4\delta^2.$$

Since $\|m_i - x\|^2 \to \delta^2$ as $i \to \infty$, we conclude that

$$\|m_j - m_i\|^2 \to 0 \qquad \text{as } i, j \to \infty.$$

Therefore, $\{m_i\}$ is a Cauchy sequence, and since M is a closed subspace of a

complete space, the sequence $\{m_i\}$ has a limit m_0 in M. By continuity of the norm, it follows that $\|x - m_0\| = \delta$. ∎

It should be noted that neither the statement nor the proof of the existence of the minimizing vector makes explicit reference to the inner product; only the norm is used. The proof, however, makes heavy use of the parallelogram law, the proof of which makes heavy use of the inner product. There is an extended version of the theorem, valid in a large class of Banach spaces, which is developed in Chapter 5, but the theorem cannot be extended to arbitrary Banach spaces.

3.4 Orthogonal Complements

In this section we apply the projection theorem to establish some additional structural properties of Hilbert space. Our primary object is to show that given any closed subspace of a Hilbert space, any vector can be written as the sum of two vectors: one in the given subspace and one orthogonal to it.

Definition. Given a subset S of a pre-Hilbert space, the set of all vectors orthogonal to S is called the *orthogonal complement* of S and is denoted $S^\perp$.

The orthogonal complement of the set consisting only of the null vector θ is the whole space. For any set S, $S^\perp$ is a closed subspace. It is a subspace because a linear combination of vectors orthogonal to a set is also orthogonal to the set. It is closed since if $\{x_n\}$ is a convergent sequence from $S^\perp$, say $x_n \to x$, continuity of the inner product implies that $0 = (x_n \mid s) \to (x \mid s)$ for all $s \in S$, so $x \in S^\perp$. The following proposition summarizes the basic relations between a set and its orthogonal complement.

Proposition 1. *Let S and T be subsets of a Hilbert space. Then*

1. *$S^\perp$ is a closed subspace.*
2. *$S \subset S^{\perp\perp}$.*
3. *If $S \subset T$, then $T^\perp \subset S^\perp$.*
4. *$S^{\perp\perp\perp} = S^\perp$.*
5. *$S^{\perp\perp} = \overline{[S]}$, i.e., $S^{\perp\perp}$ is the smallest closed subspace containing S.*

Proof.

1. This was proved in the text above.
2. If $x \in S$, then $x \perp y$ for all $y \in S^\perp$; therefore, $x \in S^{\perp\perp}$.
3. If $y \in T^\perp$, then $y \perp x$ for all $x \in S$ since $S \subset T$. Therefore, $y \in S^\perp$.
4. From 2, $S^\perp \subset S^{\perp\perp\perp}$. Also $S \subset S^{\perp\perp}$ which from 3 implies $S^{\perp\perp\perp} \subset S^\perp$. Therefore, $S^\perp = S^{\perp\perp\perp}$.
5. This statement is most easily proved by using the result of Theorem 1 below. We leave it as an exercise. ∎

Definition. We say that a vector space X is the *direct sum* of two subspaces M and N if every vector $x \in X$ has a unique representation of the form $x = m + n$ where $m \in M$ and $n \in N$. We describe this situation by the notation $X = M \oplus N$. (The only difference between direct sum and the earlier definition of sum is the added requirement of uniqueness.)

We come now to the theorem that motivates the expression "orthogonal complement" for the set of vectors orthogonal to a set. If the set is a closed subspace in a Hilbert space, its orthogonal complement contains enough additional vectors to generate the space.

Theorem 1. *If M is a closed linear subspace of a Hilbert space H, then $H = M \oplus M^{\perp}$ and $M = M^{\perp\perp}$.*

Proof. The proof follows from the projection theorem. Let $x \in H$. By the projection theorem, there is a unique vector $m_0 \in M$ such that $\|x - m_0\| \leq \|x - m\|$ for all $m \in M$, and $n_0 = x - m_0 \in M^{\perp}$. Thus, $x = m_0 + n_0$ with $m_0 \in M$ and $n_0 \in M^{\perp}$.

To show that this representation is unique, suppose $x = m_1 + n_1$ with $m_1 \in M$, $n_1 \in M^{\perp}$. Then $\theta = m_1 - m_0 + n_1 - n_0$ but $m_1 - m_0$ and $n_1 - n_0$ are orthogonal. Hence, by the Pythagorean theorem, $\|\theta\|^2 = \|m_1 - m_0\|^2 + \|n_1 - n_0\|^2$. This implies that $m_0 = m_1$ and $n_0 = n_1$.

To show that $M = M^{\perp\perp}$, it is only necessary to show $M^{\perp\perp} \subset M$ since by Proposition 1 we have $M \subset M^{\perp\perp}$. Let $x \in M^{\perp\perp}$. By the first part of this theorem, $x = m + n$ where $m \in M$, $n \in M^{\perp}$. Since both $x \in M^{\perp\perp}$ and $m \in M^{\perp\perp}$, we have $x - m \in M^{\perp\perp}$; that is, $n \in M^{\perp\perp}$. But also $n \in M^{\perp}$, hence $n \perp n$ which implies $n = \theta$. Thus, $x = m \in M$ and $M^{\perp\perp} \subset M$. ∎

In view of the above results, given a vector x and a closed subspace M in a Hilbert space, the vector $m_0 \in M$ such that $x - m_0 \in M^{\perp}$ is called the *orthogonal projection of x onto M.*

3.5 The Gram-Schmidt Procedure

Definition. A set S of vectors in a pre-Hilbert space is said to be an *orthogonal set* if $x \perp y$ for each $x, y \in S$, $x \neq y$. The set is said to be *orthonormal* if, in addition, each vector in the set has norm equal to unity.

As we shall see, orthogonal sets are extremely convenient in various problems arising in Hilbert space. Partially, this is due to the property stated below.

Proposition 1. *An orthogonal set of nonzero vectors is a linearly independent set.*

Proof. Suppose $\{x_1, x_2, \ldots, x_n\}$ is a finite subset of the given orthogonal set and that there are n scalars α_i, $i = 1, 2, \ldots, n$, such that $\sum_{i=1}^{n} \alpha_i x_i = \theta$. Taking the inner product of both sides of this equation with x_k produces

$$\left(\sum_{i=1}^{n} \alpha_i x_i \,|\, x_k\right) = (\theta \,|\, x_k)$$

or

$$\alpha_k (x_k \,|\, x_k) = 0.$$

Thus, $\alpha_k = 0$ for each k and, according to Theorem 1, Section 2.5, the vectors are independent. ∎

In Hilbert space, orthonormal sets are greatly favored over other linearly independent sets, and it is reassuring to know that they exist and can be easily constructed. The next theorem states that in fact an orthonormal set can be constructed from an arbitrary linearly independent set. The constructive method employed for the proof is referred to as the Gram-Schmidt orthogonalization procedure and is as important in its own right as the theorem itself.

Theorem 1. (*Gram-Schmidt*) *Let* $\{x_i\}$ *be a countable or finite sequence of linearly independent vectors in a pre-Hilbert space* X. *Then, there is an orthonormal sequence* $\{e_i\}$ *such that for each* n *the space generated by the first* n e_i*'s is the same as the space generated by the first* n x_i*'s; i.e., for each* n *we have* $[e_1, e_2, \ldots, e_n] = [x_1, x_2, \ldots, x_n]$.

Proof. For the first vector, take

$$e_1 = \frac{x_1}{\|x_1\|}$$

which obviously generates the same space as x_1. Form e_2 in two steps. First, put

$$z_2 = x_2 - (x_2 \,|\, e_1) e_1$$

and then $e_2 = z_2 / \|z_2\|$. By direct calculation, it is verified that $z_2 \perp e_1$ and $e_2 \perp e_1$. The vector z_2 cannot be zero since x_2 and e_1 are linearly independent; furthermore, e_2 and e_1 span the same space as x_1 and x_2 since x_2 may be expressed as a linear combination of e_1 and e_2.

The process is best understood in terms of the two-dimensional diagram of Figure 3.3. The vector z_2 is formed by subtracting the projection of x_2 on

e_1 from x_2. The remaining e_i's are defined by induction. The vector z_n is formed according to the equation

$$z_n = x_n - \sum_{i=1}^{n-1} (x_n | e_i) e_i$$

and

$$e_n = \frac{z_n}{\|z_n\|}.$$

Again it is easily verified by direct computation that $z_n \perp e_i$ for all $i < n$, and z_n is not zero since it is a linear combination of independent vectors. It is clear by induction that the e_i's generate the same space as the x_i's. If the original collection of x_i's is finite, the process terminates; otherwise the process produces an infinite sequence of orthonormal vectors. ▮

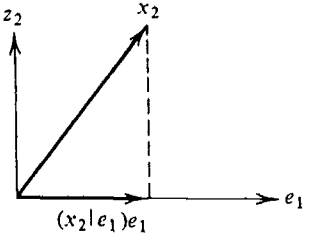

Figure 3.3 The Gram-Schmidt procedure

The study of the Gram-Schmidt procedure and its relation to the projection theorem, approximation, and equation solving is continued in the following few sections.

APPROXIMATION

3.6 The Normal Equations and Gram Matrices

In this section we investigate the following approximation problem. Suppose $y_1, y_2, \ldots, y_n$ are elements of a Hilbert space H. These vectors generate a (closed) finite-dimensional subspace M of H. Given an arbitrary vector $x \in H$, we seek the vector $\hat{x}$ in M which is closest to x. If $\hat{x}$ is expressed in terms of the vectors y_i as $\hat{x} = \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_n y_n$, the problem is equivalent to that of finding the n scalars α_i, $i = 1, 2, \ldots, n$, that minimize $\|x - \alpha_1 y_1 - \alpha_2 y_2 - \cdots - \alpha_n y_n\|$.

According to the projection theorem, the unique minimizing vector $\hat{x}$ is the orthogonal projection of x on M, or equivalently the difference vector $x - \hat{x}$ is orthogonal to each of the vectors y_i. Therefore,

$$(x - \alpha_1 y_1 - \alpha_2 y_2 - \cdots - \alpha_n y_n \mid y_i) = 0$$

for $i = 1, 2, \ldots, n$. Or, equivalently,

$$\begin{aligned}
(y_1 \mid y_1)\alpha_1 + (y_2 \mid y_1)\alpha_2 + \cdots + (y_n \mid y_1)\alpha_n &= (x \mid y_1) \\
(y_1 \mid y_2)\alpha_1 + (y_2 \mid y_2)\alpha_2 + \cdots + (y_n \mid y_2)\alpha_n &= (x \mid y_2) \\
\vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \qquad\quad \vdots & \\
(y_1 \mid y_n)\alpha_1 + \qquad \cdots \qquad + (y_n \mid y_n)\alpha_n &= (x \mid y_n).
\end{aligned}$$

These equations in the n coefficients α_i are known as the *normal equations* for the minimization problem.

Corresponding to the vectors $y_1, y_2, \ldots, y_n$, the $n \times n$ matrix

$$G = G(y_1, y_2, \ldots, y_n) = \begin{bmatrix} (y_1 \mid y_1) & (y_1 \mid y_2) & \cdots & (y_1 \mid y_n) \\ (y_2 \mid y_1) & & & \\ \vdots & & & \vdots \\ (y_n \mid y_1) & & \cdots & (y_n \mid y_n) \end{bmatrix}$$

is called the *Gram matrix* of $y_1, y_2, \ldots, y_n$. It is the transpose of the coefficient matrix of the normal equations. (In a real pre-Hilbert space the matrix is symmetric. In a complex space its transpose is its complex conjugate.) The determinant $g = g(y_1, y_2, \ldots, y_n)$ of the Gram matrix is known as the *Gram determinant*.

The approximation problem is solved once the normal equations are solved. In order that this set of equations be uniquely solvable, it is necessary and sufficient that the Gram determinant be nonzero.

Proposition 1. *$g(y_1, y_2, \ldots, y_n) \neq 0$ if and only if the vectors $y_1, y_2, \ldots, y_n$ are linearly independent.*

Proof. The equivalent contrapositive statement is that

$$g(y_1, y_2, \ldots, y_n) = 0$$

if and only if the vectors $y_1, y_2, \ldots, y_n$ are linearly dependent. Suppose that the y_i's are linearly dependent; that is, suppose there are constants α_i, not all zero, such that $\sum_{i=1}^{n} \alpha_i y_i = \theta$. It is then clear that the rows of the Gram determinant have a corresponding dependency and, hence, the determinant is zero.

Now assume that the Gram determinant is zero or, equivalently, that there is a linear dependency among the rows. In that case there are constants α_i, not all zero, such that $\sum_{i=1}^{n} \alpha_i(y_i \mid y_j) = 0$ for all j. From this it follows that

$$(\sum_{i=1}^{n} \alpha_i y_i \mid y_j) = 0 \qquad \text{for all } j$$

and hence that

$$\sum_{j=1}^{n} \bar{\alpha}_j(\sum_{i=1}^{n} \alpha_i y_i \mid y_j) = 0,$$

or

$$\left\| \sum_{i=1}^{n} \alpha_i y_i \right\|^2 = 0.$$

Thus, $\sum_{i=1}^{n} \alpha_i y_i = \theta$ and the vectors $y_1, y_2, \ldots, y_n$ are linearly dependent. ∎

Although the normal equations do not possess a unique solution if the y_i's are linearly dependent, there is always at least one solution. Thus the degeneracy that arises as a result of $g = 0$ always results in a multiplicity of solutions rather than an inconsistent set of equations. The reader is asked to verify this in Problem 4.

We turn now to the evaluation of the minimum distance between x and the subspace M.

Theorem 1. *Let $y_1, y_2, \ldots, y_n$ be linearly independent. Let δ be the minimum distance from a vector x to the subspace M generated by the y_i's; i.e.,*

$$\delta = \min \|x - \alpha_1 y_1 - \alpha_2 y_2 \ldots - \alpha_n y_n\| = \|x - \hat{x}\|.$$

Then

$$\delta^2 = \frac{g(y_1, y_2, \ldots, y_n, x)}{g(y_1, y_2, \ldots, y_n)}.$$

Proof. By definition, $\delta^2 = \|x - \hat{x}\|^2 = (x - \hat{x} \mid x) - (x - \hat{x} \mid \hat{x})$. By the projection theorem, $x - \hat{x}$ is orthogonal to M so, in particular, $(x - \hat{x} \mid \hat{x}) = 0$. Therefore,

$$\delta^2 = (x - \hat{x} \mid x) = (x \mid x) - \alpha_1(y_1 \mid x) - \alpha_2(y_2 \mid x) - \cdots - \alpha_n(y_n \mid x)$$

or

$$\alpha_1(y_1 \mid x) + \alpha_2(y_2 \mid x) + \cdots + \alpha_n(y_n \mid x) + \delta^2 = (x \mid x).$$

This equation, together with the normal equations, gives $n + 1$ linear

equations for the $n+1$ unknowns $\alpha_1, \alpha_2, \ldots, \alpha_n, \delta^2$. Applying Cramer's rule to determine δ^2, we obtain

$$\delta^2 = \frac{\begin{vmatrix} (y_1 \mid y_1) & (y_2 \mid y_1) & \cdots & (y_n \mid y_1) & (x \mid y_1) \\ (y_1 \mid y_2) & & & & \\ \vdots & & & & \vdots \\ (y_1 \mid y_n) & & \cdots & (y_n \mid y_n) & (x \mid y_n) \\ (y_1 \mid x) & & \cdots & (y_n \mid x) & (x \mid x) \end{vmatrix}}{\begin{vmatrix} (y_1 \mid y_1) & (y_2 \mid y_1) & \cdots & (y_n \mid y_1) & 0 \\ (y_1 \mid y_2) & & & & \\ \vdots & & & & \vdots \\ (y_1 \mid y_n) & & \cdots & (y_n \mid y_n) & 0 \\ (y_1 \mid x) & & \cdots & (y_n \mid x) & 1 \end{vmatrix}} = \frac{g(y_1, y_2, \ldots, y_n, x)}{g(y_1, y_2, \ldots, y_n)}. \blacksquare$$

This explicit formula, although of some theoretical interest, is of little practical importance because of the impracticality of evaluating large determinants. The evaluation of an $n+1$-dimensional determinant is in general about as difficult as inverting an n-dimensional matrix; hence, the determinant relation is not particularly attractive for direct computations. An alternative approach is developed in Section 3.9.

3.7 Fourier Series

Consider the problem of finding the best approximation to x in the subspace M generated by the orthonormal vectors $e_1, e_2, \ldots, e_n$. This special case of the general approximation problem is trivial since the Gram matrix of the e_i's is simply the identity matrix. Thus the best approximation is

$$\hat{x} = \sum_{i=1}^{n} \alpha_i e_i$$

where $\alpha_i = (x \mid e_i)$.

In this section we generalize this special approximation problem slightly by considering approximation in the closed subspace generated by an infinite orthonormal sequence. This leads us naturally to a general discussion of Fourier series.

To proceed we must define what is meant by an infinite series.

Definition. An *infinite series* of the form $\sum_{i=1}^{\infty} x_i$ is said to *converge* to the element x in a normed space if the sequence of partial sums $s_n = \sum_{i=1}^{n} x_i$ converges to x. In that case we write $x = \sum_{i=1}^{\infty} x_i$.

We now establish a necessary and sufficient condition for an infinite series of orthogonal vectors to converge in Hilbert space.

Theorem 1. *Let $\{e_i\}$ be an orthonormal sequence in a Hilbert space H. A series of the form $\sum_{i=1}^{\infty} \xi_i e_i$ converges to an element $x \in H$ if and only if $\sum_{i=1}^{\infty} |\xi_i|^2 < \infty$ and, in that case, we have $\xi_i = (x \mid e_i)$.*

Proof. Suppose that $\sum_{i=1}^{\infty} |\xi_i|^2 < \infty$ and define the sequence of partial sums $s_n = \sum_{i=1}^{n} \xi_i e_i$. Then,

$$\|s_n - s_m\|^2 = \left\| \sum_{i=n+1}^{m} \xi_i e_i \right\|^2 = \sum_{i=n+1}^{m} |\xi_i|^2 \to 0$$

as $n, m \to \infty$. Therefore, $\{s_n\}$ is a Cauchy sequence so by the completeness of H there is an element $x \in H$ such that $s_n \to x$.

Conversely, if s_n converges, it is a Cauchy sequence so $\sum_{i=n+1}^{m} |\xi_i|^2 \to 0$. Thus, $\sum_{i=n+1}^{\infty} |\xi_i|^2 \to 0$ and $\sum_{i=1}^{\infty} |\xi_i|^2 < \infty$.

Obviously, $(s_n \mid e_i) \to \xi_i$ as $n \to \infty$ which, by the continuity of the inner product (Lemma 4, Section 3.2), implies that $(x \mid e_i) = \xi_i$. ∎

In the above theorem we started with an arbitrary square-summable sequence of scalars $\{\xi_i\}$ and constructed the element $x = \sum_{i=1}^{\infty} \xi_i e_i$ in the space. It was found that $\xi_i = (x \mid e_i)$. We now consider the possibility of starting with a given x and forming an infinite summation with $(x \mid e_i)$ as coefficients of the orthonormal vectors e_i. These coefficients are called the *Fourier coefficients* of x with respect to the e_i and are fundamental in the theory of generalized Fourier series in Hilbert space.

Lemma 1. (*Bessel's Inequality*) *Let x be an element in a Hilbert space H and suppose $\{e_i\}$ is an orthonormal sequence in H. Then,*

$$\sum_{i=1}^{\infty} |(x \mid e_i)|^2 \leq \|x\|^2.$$

Proof. Letting $\alpha_i = (x \mid e_i)$, we have for all n

$$0 \leq \|x - \sum_{i=1}^{n} \alpha_i e_i\|^2 = \|x\|^2 - \sum_{i=1}^{n} |\alpha_i|^2.$$

Thus,

$$\sum_{i=1}^{n} |\alpha_i|^2 \leq \|x\|^2 \qquad \text{for all } n.$$

Hence,

$$\sum_{i=1}^{\infty} |\alpha_i|^2 \leq \|x\|^2. \quad ∎$$

Since Bessel's inequality guarantees that $\sum_{i=1}^{\infty} |(x \mid e_i)|^2 < \infty$, Theorem 1 guarantees that the series $\sum_{i=1}^{\infty} (x \mid e_i)e_i$ converges to some element. We now characterize that element.

Definition. If S is a subset of a Hilbert space, the *closed subspace generated by* S is $\overline{[S]}$.

Theorem 2. *Let x be an element in a Hilbert space H and suppose $\{e_i\}$ is a orthonormal sequence in H. Then the series*

$$\sum_{i=1}^{\infty} (x \mid e_i)e_i$$

converges to an element $\hat{x}$ in the closed subspace M generated by the e_i's. The difference vector $x - \hat{x}$ is orthogonal to M.

Proof. The fact that the series converges is a consequence of Theorem 1 and Bessel's inequality. Obviously, $\hat{x} \in M$. The sequence of partial sums $s_n = \sum_{i=1}^{n} (x \mid e_i)e_i$ converges to $\hat{x}$. For each j and for $n > j$, we have

$$(x - s_n \mid e_j) = (x - \sum_{i=1}^{n} (x \mid e_i)e_i \mid e_j) = (x \mid e_j) - (x \mid e_j) = 0.$$

Therefore, by the continuity of the inner product $\lim_{n \to \infty} (x - s_n \mid e_j) = (x - \hat{x} \mid e_j) = 0$ for each j. Thus, $x - \hat{x}$ is orthogonal to the subspace generated by the e_i's. Again using the continuity of the inner product, we deduce that $x - \hat{x}$ is perpendicular to the closed subspace generated by the e_i's. ∎

From Theorem 2 it is apparent that if the closed subspace generated by the orthonormal set of e_i's is the whole space, any vector in H can be expanded as a series of the e_i's with coefficients equal to the Fourier coefficients $(x \mid e_i)$. In the next section, we turn to the problem of constructing an orthonormal sequence that generates the whole space.

*3.8 Complete Orthonormal Sequences

Suppose $\{e_i\}$ is a sequence of orthonormal vectors in a Hilbert space H. Generally this sequence generates a linear subspace in H and, as we have seen, any vector in the closure M of this subspace can be expressed uniquely as the limit of an infinite series of the form $\sum_{i=1}^{\infty} \alpha_i e_i$. To express every vector in H this way, it is necessary that the closed subspace M generated by the e_i's be the whole space.

Definition. An orthonormal sequence $\{e_i\}$ in a Hilbert space H is said to be *complete* if the closed subspace generated by the e_i's is H.

The following result gives a simple criterion for an arbitrary orthonormal sequence to be complete. We leave the proof to the reader.

Lemma 1. *An orthonormal sequence $\{e_i\}$ in a Hilbert space H is complete if and only if the only vector orthogonal to each of the e_i's is the null vector.*

Example 1. Consider the Hilbert space $L_2[-1, 1]$ consisting of square-integrable functions on the interval $[-1, 1]$. The independent functions $1, t, t^2, \ldots, t^n, \ldots$ generate the subspace of polynomials. The Gram-Schmidt procedure can be applied to the sequence $1, t, t^2, \ldots$ to produce an orthonormal sequence $\{e_i\}_{i=0}^{\infty}$. Obviously the e_i's, being linear combinations of polynomials, are themselves polynomials. It turns out that

$$e_n(t) = \sqrt{\frac{2n+1}{2}}\, P_n(t), \qquad n = 0, 1, 2, \ldots,$$

where the $P_n(t)$ are the well-known Legendre polynomials

$$P_n(t) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dt^n} \{(1 - t^2)^n\}.$$

We wish to show that this sequence is a complete orthonormal sequence in $L_2[-1, 1]$. According to Lemma 1, it is sufficient to prove that there is no nonzero vector orthogonal to the subspace of polynomials.

Assume that there exists an $f \in L_2[-1, 1]$ orthogonal to each t^n; i.e., assume that

$$\int_{-1}^{1} t^n f(t)\, dt = 0 \qquad \text{for } n = 0, 1, 2, \ldots.$$

The continuous function

$$F(t) = \int_{-1}^{t} f(\tau)\, d\tau$$

has $F(-1) = F(1) = 0$, and it follows from integration by parts that

$$\int_{-1}^{1} t^n F(t)\, dt = \frac{t^{n+1}}{n+1} F(t)\Big|_{-1}^{1} - \int_{-1}^{1} \frac{t^{n+1}}{n+1} f(t)\, dt = 0$$

for $n = 0, 1, 2, \ldots$. Thus the continuous function F is also orthogonal to polynomials.

Since F is continuous, the Weierstrass approximation theorem applies: Given $\varepsilon > 0$, there is a polynomial

$$Q(t) = \sum_{i=0}^{N} a_i t^i$$

such that

$$|F(t) - Q(t)| < \varepsilon$$

for all $t \in [-1, 1]$. Therefore, since F is orthogonal to polynomials,

$$\int_{-1}^{1} |F(t)|^2 \, dt = \int_{-1}^{1} F(t)[F(t) - Q(t)] \, dt$$

$$\leq \varepsilon \int_{-1}^{1} |F(t)| \, dt \leq \varepsilon \sqrt{2} \left\{ \int_{-1}^{1} |F(t)|^2 \, dt \right\}^{1/2},$$

the last inequality being a special case of the Cauchy-Schwarz inequality. It follows that

$$\int_{-1}^{1} |F(t)|^2 \, dt \leq 2\varepsilon^2.$$

Since ε is arbitrary and F is continuous, we deduce that $F(t) = 0$ and, hence, that $f(t) = 0$ (almost everywhere). This proves that the subspace of polynomials is dense in $L_2[-1, 1]$.

Example 2. Consider the complex space $L_2[0, 2\pi]$. The functions $e_k(t) = (1/\sqrt{2\pi})e^{ikt}$ for $k = 0, \pm 1, \pm 2, \ldots$ are easily seen by direct calculation to be an orthonormal sequence. Furthermore, it may be shown, in a manner similar to that for the Legendre polynomials, that the system is complete. Therefore, we obtain the classical complex Fourier expansion for an arbitrary function $x \in L_2[0, 2\pi]$

$$x(t) = \sum_{k=-\infty}^{\infty} f_k \frac{e^{ikt}}{\sqrt{2\pi}}$$

where the Fourier coefficients c_k are evaluated according to the formula

$$c_k = (x \mid e_k) = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} x(t) e^{-ikt} \, dt.$$

3.9 Approximation and Fourier Series

Suppose again that we are given independent vectors $y_1, y_2, \ldots, y_n$ generating a subspace M of a Hilbert space H and wish to find the vector $\hat{x}$ in M which minimizes $\|x - \hat{x}\|$. Rather than seeking to obtain $\hat{x}$ directly as a linear combination of the y_i's by solving the normal equations, we can employ the Gram-Schmidt orthogonalization procedure together with Fourier series.

First we apply the Gram-Schmidt procedure to $\{y_1, y_2, \ldots, y_n\}$, obtaining an orthonormal set $\{e_1, e_2, \ldots, e_n\}$ generating M. The vector $\hat{x}$ is then given by the Fourier series

$$\hat{x} = \sum_{i=1}^{n} (x \mid e_i) e_i$$

since $x - \hat{x}$ is orthogonal to M. Thus, our original optimization problem is easily resolved once the independent vectors y_i are orthonormalized. The advantage of this method is that once the e_i's are found, the best approximation to any vector is easily computed.

Since solution to the approximation problem is equivalent to solution of the normal equations, it is clear that the Gram-Schmidt procedure can be interpreted as a procedure for inverting the Gram matrix. Conversely, many effective algorithms for solving the normal equations have an interpretation in terms of the minimum norm problem in Hilbert space.

We have seen that the Gram-Schmidt procedure can be used to solve a minimum norm approximation problem. It is interesting to note that the Gram-Schmidt procedure can itself be viewed as an approximation problem. Given a sequence $\{y_1, y_2, y_3, \ldots, y_n\}$ of independent vectors, the Gram-Schmidt procedure sets

$$e_1 = \frac{y_1}{\|y_1\|}$$

$$e_k = \frac{y_k - \sum_{i=1}^{k-1} (y_k \mid e_i) e_i}{\left\| y_k - \sum_{i=1}^{k-1} (y_k \mid e_i) e_i \right\|}.$$

The vector $y_k - \sum_{i=1}^{k-1} (y_k \mid e_i) e_i$ is the optimal error for the minimum problem:

$$\text{minimize } \|y_k - \hat{x}\|$$

where the minimum is over all $\hat{x}$ in the subspace $[y_1, y_2, \ldots, y_{k-1}] = [e_1, e_2, \ldots, e_{k-1}]$. The vector e_k is just a normalized version of this error. Thus, the Gram-Schmidt procedure consists of solving a series of minimum norm approximation problems by use of the projection theorem.

Alternatively, the minimum norm approximation of x on the subspace $[y_1, y_2, \ldots, y_n]$ can be found by applying the Gram-Schmidt procedure to the sequence $\{y_1, y_2, y_3, \ldots, y_n, x\}$. The optimal error $x - \hat{x}$ is found at the last step.

OTHER MINIMUM NORM PROBLEMS

3.10 The Dual Approximation Problem

In Section 3.6 we considered in some detail the problem of approximating an arbitrary vector in a Hilbert space by a vector in a given finite-dimensional subspace. The projection theorem led to the normal equations which could be solved for the best approximation. A major assumption in such problems was the finite dimensionality of the subspace from which the approximation was chosen. Finite dimensionality not only guarantees closure (and hence existence of a solution) but leads to a feasible computation procedure for obtaining the solution.

In many important and interesting practical problems the subspace in which the solution must lie is not finite dimensional. In such problems it is generally not possible to reduce the problem to a finite set of linear equations. However, there is an important class of such problems that can be reduced by the projection theorem to a finite set of linear equations similar to the normal equations. In this section we study these problems and their relation to the earlier approximation problem. We begin by pointing out a trivial modification of the projection theorem applicable to linear varieties.

Theorem 1. (*Restatement of Projection Theorem*) *Let M be a closed subspace of a Hilbert space H. Let x be a fixed element in H and let V be the linear variety $x + M$. Then there is a unique vector x_0 in V of minimum norm. Furthermore, x_0 is orthogonal to M.*

Proof. The theorem is proved by translating V by $-x$ so that it becomes a closed subspace and then applying the projection theorem. See Figure 3.4. ∎

A point of caution is necessary here. The minimum norm solution x_0 is *not* orthogonal to the linear variety V but to the subspace M from which V is obtained.

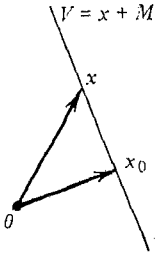

Figure 3.4 Minimum norm to a linear variety

Two special kinds of linear varieties are of particular interest in optimization theory because they lead to finite-dimensional problems. The first is the n-dimensional linear variety consisting of points of the form $x + \sum_{i=1}^{n} a_i x_i$ where $\{x_1, x_2, \ldots, x_n\}$ is a linearly independent set in H and x is a fixed vector in H. Problems which seek minimum norm vectors in an n-dimensional variety can be reduced to the solution of an n-dimensional set of normal equations as developed in Section 3.6.

The second special type of linear variety consists of all vectors x in a Hilbert space H satisfying conditions of the form

$$\begin{aligned}(x \mid y_1) &= c_1\\ (x \mid y_2) &= c_2\\ &\vdots\\ (x \mid y_n) &= c_n\end{aligned}$$

where $y_1, y_2, \ldots, y_n$ are a set of linearly independent vectors in H and $c_1, c_2, \ldots, c_n$ are fixed constants. If we denote by M the subspace generated by $y_1, y_2, \ldots, y_n$, it is clear that if each $c_i = 0$, then the linear variety is the subspace $M^{\perp}$. For nonzero c_i's the resulting linear variety is a translation of $M^{\perp}$. A linear variety of this form is said to be of *codimension n* since the orthogonal complement of the subspace producing it has dimension n.

We now consider the minimum norm problem of seeking the closest vector to the origin lying in a linear variety of finite codimension.

Theorem 2. *Let H be a Hilbert space and $\{y_1, y_2, \ldots, y_n\}$ a set of linearly independent vectors in H. Among all vectors $x \in H$ satisfying*

$$\begin{aligned}(x \mid y_1) &= c_1\\ (x \mid y_2) &= c_2\\ &\vdots\\ (x \mid y_n) &= c_n,\end{aligned}$$

let x_0 have the minimum norm. Then

$$x_0 = \sum_{i=1}^{n} \beta_i y_i$$

where the coefficients β_i satisfy the equations

$$\begin{aligned}(y_1 \mid y_1)\beta_1 + (y_2 \mid y_1)\beta_2 + \cdots + (y_n \mid y_1)\beta_n &= c_1\\ (y_1 \mid y_2)\beta_1 + (y_2 \mid y_2)\beta_2 + \cdots + (y_n \mid y_2)\beta_n &= c_2\\ &\vdots\\ (y_1 \mid y_n)\beta_1 + \cdots + (y_n \mid y_n)\beta_n &= c_n.\end{aligned}$$

Proof. Let M be the (n-dimensional) subspace generated by the vectors y_i. The linear variety defined by the constraints of the minimization problem is a translation of the subspace $M^\perp$. Since $M^\perp$ is closed, existence and uniqueness of an optimal solution follow from the modified projection theorem. Furthermore, the optimal solution x_0 is orthogonal to $M^\perp$. Thus, $x_0 \in M^{\perp\perp}$. However, since M is closed, $M^{\perp\perp} = M$. We therefore conclude that $x_0 \in M$ or, equivalently, that

$$x_0 = \sum_{i=1}^{n} \beta_i y_i.$$

The n coefficients β_i are chosen so that x_0 satisfies the n constraints. This leads to the n equations displayed in the theorem statement. ▮

Example 1. The shaft angular velocity ω of a d-c motor driven from a variable current source u is governed by the following first-order differential equation:

$$\dot{\omega}(t) + \omega(t) = u(t),$$

where $u(t)$ is the field current at time t. The angular position θ of the motor shaft is the time integral of ω. Assume that the motor is initially at rest, $\theta(0) = \omega(0) = 0$, and that it is desired to find the field current function u of minimum energy which rotates the shaft to the new rest position $\theta = 1$, $\omega = 0$ within one second. The energy is assumed to be proportional to $\int_0^1 u^2(t)\,dt$. This is a simple control problem in which the cost criterion depends only on the control function u. The problem can be solved by treating it as a minimum norm problem in the Hilbert space $L_2[0, 1]$.

We can integrate the first-order differential equation governing the motor to obtain the explicit formula

$$\omega(1) = \int_0^1 e^{(t-1)} u(t)\,dt \tag{1}$$

for the final angular velocity corresponding to any control. From the equation $\dot{\omega}(t) + \dot{\theta}(t) = u(t)$ follows the equation

$$\theta(1) = \int_0^1 u(t)\,dt - \omega(1) \tag{2}$$

or, combined with equation (1),

$$\theta(1) = \int_0^1 \{1 - e^{(t-1)}\} u(t)\,dt. \tag{3}$$

If the function u is regarded as an element of the Hilbert space $L_2[0, 1]$, the above relations may be expressed as

$$\omega(1) = (y_1 \mid u)$$
$$\theta(1) = (y_2 \mid u)$$

where $y_1 = e^{(t-1)}$, $y_2 = 1 - e^{(t-1)}$.

With these definitions the original problem is equivalent to that of finding $u \in L_2[0, 1]$ of minimum norm subject to the constraints

$$\begin{aligned} 0 &= (y_1 \mid u) \\ 1 &= (y_2 \mid u). \end{aligned} \tag{4}$$

According to Theorem 2, the optimal solution is in the subspace generated by y_1 and y_2. Equivalently,

$$u(t) = \alpha_1 + \alpha_2 e^t$$

where α_1 and α_2 are chosen to satisfy the two constraints (4). Evaluating the constants leads to the solution

$$u(t) = \frac{1}{3 - e}[1 + e - 2e^t].$$

We have observed that there are two basic forms of minimum norm problems in Hilbert space that reduce to the solution of a finite number of simultaneous linear equations. In their general forms, both problems are concerned with finding the shortest distance from a point to a linear variety. In one case the linear variety has finite dimension; in the other it has finite codimension. To understand the relation between these two problems, which is one of the simplest examples of a duality relation, consider Figure 3.5. Let M be a closed subspace of a Hilbert space H and let x be an arbitrary vector in H. We may then formulate two problems: one of projecting x onto M and the other of projecting x onto $M^\perp$. The situation is completely symmetrical since $M^{\perp\perp} = M$. If we solve one of these problems, the other is automatically solved in the process since if, for instance,

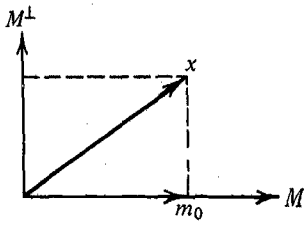

Figure 3.5 Dual projection problems

m_0 is the projection of x onto M, then $x - m_0$ is the projection of x onto $M^\perp$. Therefore, if either M or $M^\perp$ is finite dimensional, both problems can be reduced to that of solving a finite set of linear equations.

*3.11 A Control Problem

In this section we indicate how the projection theorem can be applied to more complicated problems to establish existence and uniqueness results. We consider an optimal control problem that is more difficult than the one discussed in the previous section but which is of great practical importance.

Suppose we seek to minimize the quadratic objective

$$J = \int_0^T \{x^2(t) + u^2(t)\}\, dt$$

where x and u are connected by the differential equation

$$\dot{x}(t) = u(t); \qquad x(0) \text{ given.} \tag{1}$$

Problems of this form arise when it is desired to reduce x to zero quickly by suitable application of control u. The quadratic objective represents a common compromise between a desire to have x small while simultaneously conserving control energy.

For notational ease we have selected the simplest possible differential equation to represent the controlled systems, but the techniques developed below apply equally well to more complex differential equations.

Let us begin by replacing equation (1) by the equivalent constraint

$$x(t) = x(0) + \int_0^t u(\tau)\, d\tau. \tag{2}$$

We now may formulate our problem in the Hilbert space

$$H = L_2[0, T] \times L_2[0, T]$$

consisting of ordered pairs (x, u) of square-integrable functions on $[0, T]$ with inner product defined by

$$((x_1, u_1) | (x_2, u_2)) = \int_0^T [x_1(t)x_2(t) + u_1(t)u_2(t)]\, dt$$

and the corresponding norm

$$\|(x, u)\|^2 = \int_0^T [x^2(t) + u^2(t)]\, dt.$$

The set of elements $(x, u) \in H$ satisfying the constraint (2) is a linear variety V in H. Thus, abstractly the control problem is one of finding the element $(x, u) \in V$ having minimum norm.

To establish the existence and uniqueness of a solution, it is sufficient to prove that the linear variety V is closed. For this purpose, let $\{(x_n, u_n)\}$ be a sequence of elements from V converging to an element (x, u). To prove that V is closed, we must show that $(x, u) \in V$. Letting $y(t) = x(0) + \int_0^t u(\tau)\, d\tau$, we must show that $x = y$. We have $y(t) - x_n(t) = \int_0^t [u(\tau) - u_n(\tau)]\, d\tau$. Thus, by the Cauchy-Schwarz inequality (applied to the functions 1 and $u - u_n$),

$$|y(t) - x_n(t)|^2 \leq t \int_0^t |u(\tau) - u_n(\tau)|^2\, d\tau \leq T\|u - u_n\|^2$$

and, hence, integrating from 0 to T, $\|y - x_n\| \leq T\|u - u_n\|$. It follows that

$$\|y - x\| \leq \|y - x_n\| + \|x_n - x\| \leq T\|u - u_n\| + \|x_n - x\|.$$

Both terms on the right tend to zero as $n \to \infty$, from which we conclude that $x = y$, the desired result.

The general quadratic loss control problem is treated further in Section 9.5.

3.12 Minimum Distance to a Convex Set

Much of the discussion in the previous sections can be generalized from linear varieties to convex sets. The following theorem, which treats the minimum norm problem illustrated in Figure 3.6, is a direct extension of the proof of the projection theorem.

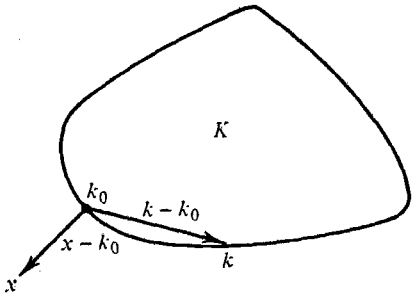

Figure 3.6 Minimum distance to a convex set

Theorem 1. *Let x be a vector in a Hilbert space H and let K be a closed convex subset of H. Then there is a unique vector $k_0 \in K$ such that*

$$\|x - k_0\| \leq \|x - k\|$$

for all $k \in K$. Furthermore, a necessary and sufficient condition that k_0 be the unique minimizing vector is that $(x - k_0 \mid k - k_0) \leq 0$ for all $k \in K$.

Proof. To prove existence, let $\{k_i\}$ be a sequence from K such that

$$\|x - k_i\| \to \delta = \inf_{k \in K} \|x - k\|.$$

By the parallelogram law,

$$\|k_i - k_j\|^2 = 2\,\|k_i - x\|^2 + 2\,\|k_j - x\|^2 - 4\left\|x - \frac{k_i + k_j}{2}\right\|^2.$$

By convexity of K, $(k_i + k_j)/2$ is in K; hence,

$$\left\|x - \frac{k_i + k_j}{2}\right\| \geq \delta$$

and therefore

$$\|k_i - k_j\|^2 \leq 2\,\|k_i - x\|^2 + 2\,\|k_j - x\|^2 - 4\delta^2 \to 0.$$

The sequence $\{k_i\}$ is Cauchy and hence convergent to an element $k_0 \in K$. By continuity, $\|x - k_0\| = \delta$.

To prove uniqueness, suppose $k_1 \in K$ with $\|x - k_1\| = \delta$. The sequence

$$k_n = \begin{cases} k_0 & n \text{ even} \\ k_1 & n \text{ odd} \end{cases}$$

has $\|x - k_n\| \to \delta$ so, by the above argument, $\{k_n\}$ is Cauchy and convergent. This can only happen if $k_1 = k_0$.

We show now that if k_0 is the unique minimizing vector, then

$$(x - k_0 \,|\, k - k_0) \leq 0$$

for all $k \in K$. Suppose to the contrary that there is a vector $k_1 \in K$ such that $(x - k_0 \,|\, k_1 - k_0) = \varepsilon > 0$. Consider the vectors $k_\alpha = (1 - \alpha)k_0 + \alpha k_1$; $0 \leq \alpha \leq 1$. Since K is convex, each $k_\alpha \in K$. Also

$$\begin{aligned}\|x - k_\alpha\|^2 &= \|(1 - \alpha)(x - k_0) + \alpha(x - k_1)\|^2 \\ &= (1 - \alpha)^2\|x - k_0\|^2 + 2\alpha(1 - \alpha)(x - k_0 \,|\, x - k_1) + \alpha^2\|x - k_1\|^2.\end{aligned}$$

The quantity $\|x - k_\alpha\|^2$ is a differentiable function of α with derivative at $\alpha = 0$ equal to

$$\begin{aligned}\frac{d}{d\alpha}\|x - k_\alpha\|^2\Big|_{\alpha=0} &= -2\,\|x - k_0\|^2 + 2(x - k_0 \,|\, x - k_1) \\ &= -2(x - k_0 \,|\, k_1 - k_0) = -2\varepsilon < 0.\end{aligned}$$

Thus for some small positive α, $\|x - k_\alpha\| < \|x - k_0\|$ which contradicts the minimizing property of k_0. Hence, no such k_1 can exist.

Conversely, suppose that $k_0 \in K$ is such that $(x - k_0 \mid k - k_0) \leq 0$ for all $k \in K$. Then for any $k \in K$, $k \neq k_0$, we have

$$\|x - k\|^2 = \|x - k_0 + k_0 - k\|^2 = \|x - k_0\|^2 + 2(x - k_0 \mid k_0 - k) + \|k_0 - k\|^2 > \|x - k_0\|^2$$

and therefore k_0 is a unique minimizing vector. ∎

Example 1. As an application of the above result, we consider an approximation problem with restrictions on the coefficients. Let $\{y_1, y_2, \ldots, y_n\}$ be linearly independent vectors in a Hilbert space H. Given $x \in H$, we seek to minimize $\|x - \alpha_1 y_1 - \alpha_2 y_2 - \cdots - \alpha_n y_n\|$ where we require $\alpha_i \geq 0$ for each i. Such a restriction is common in many physical problems.

This general problem can be formulated abstractly as that of finding the minimum distance from the point x to the convex cone

$$K = \{y : y = \alpha_1 y_1 + \cdots + \alpha_n y_n, \alpha_i \geq 0 \text{ each } i\}.$$

This cone is obviously closed and there is therefore a unique minimizing vector. The minimizing vector $\hat{x} = \alpha_1 y_1 + \cdots + \alpha_n y_n$ must satisfy

$$(x - \hat{x} \mid k - \hat{x}) \leq 0, \qquad \text{all } k \in K.$$

Setting $k = \hat{x} + y_i$ leads to

$$(x - \hat{x} \mid y_i) \leq 0 \qquad \text{for } i = 1, 2, \ldots, n$$

and setting $k = \hat{x} - \alpha_i y_i$ leads to

$$(x - \hat{x} \mid y_i) \geq 0 \qquad \text{if } \alpha_i > 0.$$

Thus it follows that

$$(x - \hat{x} \mid y_i) \leq 0 \qquad \text{for } i = 1, 2, \ldots, n$$

with equality if $\alpha_i > 0$.

Letting G be the Gram matrix of the y_i's and letting $b_i = (x \mid y_i)$, we obtain the equation

$$G\alpha - b = z \tag{1}$$

for some vector z with components $z_i \geq 0$. (Here α and b are vectors with components α_i and b_i, respectively.) Furthermore, $\alpha_i z_i = 0$ for each i or, more compactly,

$$\alpha' z = 0. \tag{2}$$

Conditions (1) and (2) are the analog of the normal equations. They represent necessary and sufficient conditions for the solution of the approximation problem but are not easily solved since they are nonlinear.

The dual of the approximation problem discussed above is that of minimizing $\|x\|$ when x is subject to inequality constraints of the form $(x \mid y_i) \geq c_i$. This topic is treated in the problems and in Chapter 5.

3.13 Problems

1. Let x and y be vectors in a pre-Hilbert space. Show that $|(x \mid y)| = \|x\| \|y\|$ if and only if $\alpha x + \beta y = 0$ for some scalars α, β.
2. Consider the set X of all real functions x defined on $(-\infty, \infty)$ for which

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} |x(t)|^2 \, dt < \infty.$$

It is easily verified that these form a linear vector space. Let M be the subspace on which the indicated limit is zero.

(a) Show that the space $H = X/M$ becomes a pre-Hilbert space when the inner product is defined as

$$([x] \mid [y]) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t) y(t) \, dt$$

(b) Show that H is not separable.

3. Let H consist of all $m \times n$ real matrices with addition and scalar multiplication defined as the usual corresponding operations with matrices, and with the inner product of two matrices A, B defined as

$$(A \mid B) = \text{Trace } [A'QB]$$

where A' denotes the transpose of the matrix A and Q is a symmetric, positive-definite $m \times m$ matrix. Prove that H is a Hilbert space.

4. Show that if $g(x_1, x_2, \ldots, x_n) = 0$, the normal equations possess a solution but it is not unique.
5. Find the linear function $x(t) = a + bt$ minimizing $\int_{-1}^{1} [t^2 - x(t)]^2 \, dt$.
6. Given a function $x \in L_2[0, 1]$, we seek a polynomial p of degree n or less which minimizes $\int_0^1 |x(t) - p(t)|^2 \, dt$ while satisfying the requirement $\int_0^1 p(t) \, dt = 0$.

(a) Show that this problem has a unique solution.

(b) Show that this problem can be solved by first finding the polynomial q of degree n or less which minimizes $\int_0^1 |x(t) - q(t)|^2 \, dt$ and then finding p of degree n or less which minimizes $\int_0^1 |q(t) - p(t)|^2 \, dt$ while satisfying $\int_0^1 p(t) \, dt = 0$.

7. Let M and N be orthogonal closed subspaces of a Hilbert space H and let x be an arbitrary vector in H. Show that the subspace $M \oplus N$ is

closed and that the orthogonal projection of x onto $M \oplus N$ is equal to $m + n$, where m is the orthogonal projection of x onto M and n is the orthogonal projection of x onto N.

8. Let $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$ each be sets of linearly independent vectors in a Hilbert space generating the subspaces M and N, respectively. Given a vector $x \in H$, it is desired to find the best approximation $\hat{x}$ (in the sense of the norm) to x in the subspace $M \cap N$.
   (a) Give a set of equations analogous to the normal equations which when solved produce $\hat{x}$.
   (b) Give a geometrical interpretation of the method of solution.
   (c) Give a computational procedure for producing $\hat{x}$.
   Hint: Consider the dual problem and use Problem 7.
9. Prove part 5 of Proposition 1, Section 3.4.
10. A Hilbert space H of functions on a set S is said to be a *reproducing kernel Hilbert space* if there is a function K defined on $S \times S$ having the properties: (1) $K(\cdot, t) \in H$ for each $t \in S$, and (2) $x(t) = (x \mid K(\cdot, t))$ for each $x \in H$, $t \in S$. Such a function K is called a reproducing kernel. Prove that a reproducing kernel, if it exists, is unique.
11. Suppose H with reproducing kernel K is a closed subspace of a Hilbert space X. Show that for $x \in X$ the function $(x \mid K(\cdot, t))$ is the projection of x onto H.
12. Suppose randomly varying data in the form of a function $x(t)$ is observed from $t = 0$ to the present time $t = T$. One proposed method for predicting the future data ($t > T$) is by fitting an $(n-1)$-th degree polynomial to the past data and using the extrapolated values of the polynomial as the estimate.

    Specifically, suppose that the polynomial,

$$p(T, t) = a_1(T) + a_2(T)t + a_3(T)t^2 + \cdots + a_n(T)t^{n-1}$$

    minimizes

$$\int_0^T [x(t) - p(T, t)]^2 \, dt.$$

    Show that the coefficients $a_i(T)$ need not be completely recomputed for each T but rather can be continuously updated according to a formula of the form

$$\frac{d}{dT} a_i(T) = \frac{b_i \varepsilon(T)}{T^i}$$

    where $\varepsilon(T) = x(T) - p(T, T)$ is the instantaneous error and the b_i's are fixed constants.

13. Show that the Gram determinant $g(x_1, x_2, \ldots, x_n)$ is never negative (thus generalizing the Cauchy-Schwarz inequality which can be expressed as $g(x_1, x_2) \geq 0$).
14. Let $\{y_1, y_2, \ldots, y_n\}$ be linearly independent vectors in a pre-Hilbert space X and let x be an arbitrary element of X. Show that the best approximation to x in the subspace generated by the y_i's has the explicit representation

$$\hat{x} = \frac{\begin{vmatrix} (y_1 \mid y_1) & (y_2 \mid y_1) & \cdots & (y_n \mid y_1) & (x \mid y_1) \\ (y_1 \mid y_2) & (y_2 \mid y_2) & \cdots & (y_n \mid y_2) & (x \mid y_2) \\ \vdots & \vdots & & \vdots & \vdots \\ (y_1 \mid y_n) & (y_2 \mid y_n) & \cdots & (y_n \mid y_n) & (x \mid y_n) \\ y_1 & y_2 & \cdots & y_n & \theta \end{vmatrix}}{-g(y_1, y_2, \ldots, y_n)}$$

where the determinant in the numerator is to be expanded algebraically to yield a linear combination of the y_i's. Show that the minimum error $\hat{x} - x$ is given by the identical formula except for x replacing θ in the determinant.

15. (Müntz's Theorem) It was shown in Example 1, Section 3.8, that the functions $1, t, t^2, \ldots$ generate a dense subspace of $L_2[-1, 1]$ (or $L_2[0, 1]$ by a simple translation and scalar factor). In this problem we prove that the functions $t^{n_1}, t^{n_2}, \ldots$; $1 \leq n_1 < n_2 < \cdots$, generate a dense subspace of L_2 [0, 1] if and only if the integer's n_i satisfy

$$\sum_{i=1}^{\infty} \frac{1}{n_i} = \infty.$$

(a) Let $M_k = [t^{n_1}, t^{n_2}, \ldots, t^{n_k}]$. The result holds if and only if for each $m \geq 1$ the minimal distance d_k of t^m from M_k goes to zero as k goes to infinity. This is equivalent to

$$\lim_{k \to \infty} \frac{g(t^{n_1}, t^{n_2}, \ldots, t^{n_k}, t^m)}{g(t^{n_1}, t^{n_2}, \ldots, t^{n_k})} = 0.$$

Show that

$$d_k^2 = \frac{\prod_{i=1}^{k} (n_i - m)^2}{(2m+1) \prod_{i=1}^{k} (m + n_i + 1)^2}.$$

(b) Show that a series of the form $\sum_{i=1}^{\infty} \log (1 + a_i)$ diverges if and only if $\sum_{i=1}^{\infty} a_i$ diverges.

(c) Show that $\lim_{k \to \infty} \log d_k^2 = -\infty$ if and only if $\sum_{i=1}^{\infty} 1/n_i = \infty$.

16. Prove Parseval's equality: An orthonormal sequence $\{e_i\}_{i=1}^{\infty}$ is complete in a Hilbert space H if and only if for each x, y in H

$$(x \mid y) = \sum_{i=1}^{\infty} (x \mid e_i)(e_i \mid y).$$

17. Let $\{y_1, y_2, \ldots, y_n\}$ be independent and suppose $\{e_1, e_2, \ldots, e_n\}$ are obtained from the y_i's by the Gram-Schmidt procedure. Let

$$\hat{x} = \sum_{i=1}^{n} (x \mid e_i)e_i = \sum_{i=1}^{n} \alpha_i y_i.$$

Show that the coefficients α_i can be easily obtained from the Fourier coefficients $(x \mid e_i)$.

18. Let $w(t)$ be a positive (weight) function defined on an interval $[a, b]$ of the real line. Assume that the integrals

$$\int_a^b t^n w(t)\, dt \qquad \text{exist for } n = 1, 2, \ldots.$$

Define the inner product of two polynomials p and q as

$$(p \mid q) = \int_a^b p(t)q(t)w(t)\, dt.$$

Beginning with the sequence $\{1, t, t^2, \ldots\}$, we can employ the Gram-Schmidt procedure to produce a sequence of orthonormal polynomials with respect to this weight function. Show that the zeros of the real orthonormal polynomials are real, simple, and located on the interior of $[a, b]$.

19. A sequence of orthonormal polynomials $\{e_n\}$ (with respect to a weighting function on a given interval) can be generated by applying the Gram-Schmidt procedure to the vectors $\{1, t, t^2, \ldots\}$. The procedure is straightforward but becomes progressively more complex with each new term. A superior technique, especially suited to machine computation, is to exploit a recursive relation of the form

$$e_n(t) = (a_n t + b_n)e_{n-1}(t) - c_n e_{n-2}(t), \qquad n = 2, 3, \ldots.$$

Show that such a recursive relation exists among the orthonormal polynomials and determine the coefficients a_n, b_n, c_n.

20. Suppose we are to set up a special manufacturing company which will operate for only ten months. During the ten months the company is to produce one million copies of a single product. We assume that the manufacturing facilities have been leased for the ten-month period, but that labor has not yet been hired. Presumably, employees will be hired

and fired during the ten-month period. Our problem is to determine how many employees should be hired or fired in each of the ten months.

It is assumed that each employee produces one hundred items per month. The cost of labor is proportional to the number of productive workers, but there is an additional cost due to hiring and firing. If $u(k)$ workers are hired in the k-th month (negative $u(k)$ corresponds to firings), the processing cost can be argued to be $u^2(k)$ because, as u increases, people must be paid to stand in line and more nonproductive employees must be paid.

At the end of the ten-month period all workers must be fired. Find $u(k)$ for $k = 1, 2, \ldots, 10$.

21. Using the projection theorem, solve the finite-dimensional problem:

$$\begin{aligned} &\text{minimize } x'Qx \\ &\text{subject to } Ax = b \end{aligned}$$

where x is an n-vector, Q a positive-definite symmetric matrix, A an $m \times n$ matrix ($m < n$), and b an m-vector.

22. Let x be a vector in a Hilbert space H, and suppose $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$ are sets of linearly independent vectors in H. We seek the vector $\hat{x}$ minimizing $\|x - \hat{x}\|$ while satisfying:

$$\hat{x} \in M = [x_1, x_2, \ldots, x_n]$$

$$(\hat{x} \mid y_i) = c_i, \qquad i = 1, 2, \ldots, m.$$

(a) Find equations for the solution which are similar to the normal equations.

(b) Give a geometric interpretation of the equations.

23. Consider the problem of finding the vector x of minimum norm satisfying

$$(x \mid y_i) \geq c_i, \qquad i = 1, 2, \ldots, n$$

where the y_i's are linearly independent.

(a) Show that this problem has a unique solution.

(b) Show that a necessary and sufficient condition that

$$x = \sum_{i=1}^{n} a_i y_i$$

be the solution is that the vector a with components a_i satisfy

$$G'a \geq c$$

$$a \geq \theta$$

and that $a_i = 0$ if $(x \mid y_i) > c_i$. G is the Gram matrix of $\{y_1, y_2, \ldots, y_n\}$.

24. The following theorem is valid in a Hilbert space H. If K is a closed convex set in H and $x \in H$, $x \notin K$; there is a unique vector $k_0 \in K$ such that $\|x - k_0\| \leq \|x - k\|$ all $k \in K$. Show that this theorem does not apply in arbitrary Banach space.

REFERENCES

§3.1–3.8. Most of the texts on functional analysis listed as references for Chapter 2 contain discussions of pre-Hilbert space and Hilbert space. In addition, see Berberian [21], or Akhiezer and Glazman [3] for introductory treatments.

§3.9. For a discussion of Gram matrices and polynomial approximation, see Davis [36] or Natanson [109].

§3.10. For some additional applications of the dual problem, see Davis [36].

§3.11. For a different proof of the existence and uniqueness to this type of control problem, see Bellman, Glicksberg, and Gross [17].

§3.13. For Problem 12, see Luenberger [99]. Reproducing kernel Hilbert space introduced in Problems 10 and 11 is a useful concept in approximation and estimation problems. See Aronszajn [11] and Parzen [116]. For additional material on polynomial approximation and a treatment of Problems 14, 15, 18, and 19, see Davis [36].

4
LEAST-SQUARES ESTIMATION

4.1 Introduction

Perhaps the richest and most exciting area of application of the projection theorem is the area of statistical estimation. It appears in virtually all branches of science, engineering, and social science for the analysis of experimental data, for control of systems subject to random disturbances, or for decision making based on incomplete information.

All estimation problems discussed in this chapter are ultimately formulated as equivalent minimum norm problems in Hilbert space and are resolved by an appropriate application of the projection theorem. This approach has several practical advantages but limits our estimation criteria to various forms of least squares. At the outset, however, it should be pointed out that there are a number of different least-squares estimation procedures which as a group offer broad flexibility in problem formulation. The differences lie primarily in the choice of optimality criterion and in the statistical assumptions required. In this chapter three basic forms of least-squares estimation are distinguished and examined.

Least squares is, of course, only one of several established approaches to estimation theory, the main alternatives being maximum likelihood and Bayesian techniques. These other techniques usually require a complete statistical description of the problem variables in terms of joint probability distribution functions, whereas least squares requires only means, variances, and covariances. Although a thorough study of estimation theory would certainly include other approaches as well as least squares, we limit our discussion to those techniques that are derived as applications of the projection theorem. In complicated, multivariable problems the equations resulting from the other approaches are often nonlinear, difficult to solve, and impractical to implement. It is only when all variables have Gaussian statistics that these techniques produce linear equations, in which case the estimate is identical with that obtained by least squares. In many practical situations then, the analyst is forced by the complexity of the problem to either assume Gaussian statistics or to employ a least-squares approach. Since the resulting estimates are identical, which is used is primarily a matter of taste.

The first few sections of this chapter are devoted to constructing various Hilbert spaces of random variables and to finding the appropriate normal equations for three basic estimation problems. No new optimization techniques are involved, but these problems provide interesting applications of the projection theorem.

4.2 Hilbert Space of Random Variables

A complete and rigorous definition of a random variable and associated probabilistic concepts is in the realm of measure and integration theory. The topic of least-squares estimation, however, makes only casual, indirect reference to the machinery available for a thorough study of probability. Consequently, the notions of probability measure and related concepts can be largely suppressed from the studies of this chapter. We assume only a rudimentary familiarity with random variables, probability distributions, and expected value. Of these, only the concept of expected value is used explicitly in least-squares estimation. With this in mind, we now briefly review those probabilistic concepts pertinent to our development.

If x is a real-valued random variable, we define the *probability distribution* P of x by

$$P(\xi) = \text{Prob}\,(x \leq \xi).$$

In other words, $P(\xi)$ is the probability that the random variable x assumes a value less than or equal to the number ξ. The *expected value* of any function g of x is then defined as

$$E[g(x)] = \int_{-\infty}^{\infty} g(\xi)\, dP(\xi),$$

which may in general not be finite. Of primary interest are the quantities

$E(x)$,	the *expected value of* x,
$E(x^2)$,	the *second moment of* x,
$E[(x - E(x))^2]$,	the *variance of* x.

Given a finite collection of real random variables $\{x_1, x_2, \ldots, x_n\}$, we define their *joint probability distribution* P as

$$P(\xi_1, \xi_2, \ldots, \xi_n) = \text{Prob}\,(x_1 \leq \xi_1, x_2 \leq \xi_2, \ldots, x_n \leq \xi_n),$$

i.e., the probability of the simultaneous occurrence of $x_i \leq \xi_i$ for all i. The expected value of any function g of the x_i's is defined as

$$E[g(x_1, x_2, \ldots, x_n)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\xi_1, \xi_2, \ldots, \xi_n)\, dP(\xi_1, \xi_2, \ldots, \xi_n).$$

All of the second-order statistical averages of these variables are described by the n expected values

$$E(x_i), \qquad i = 1, 2, \ldots, n,$$

and the $n \times n$ *covariance matrix* cov $(x_1, x_2, \ldots, x_n)$ whose ij-th element is defined as

$$E\{[x_i - E(x_i)][x_j - E(x_j)]\}$$

which in case of zero means reduces to $E(x_i x_j)$.

Two random variables x_i, x_j are said to be *uncorrelated* if $E(x_i x_j) = E(x_i)E(x_j)$ or, equivalently, if the ij-th element of the covariance matrix is zero.

With this elementary background material, we now construct a Hilbert space of random variables. Let $\{y_1, y_2, \ldots, y_m\}$ be a finite collection of random variables with $E(y_i^2) < \infty$ for each i. We define a Hilbert space H consisting of all random variables that are linear combinations of the y_i's. The inner product of two elements x, y in H is defined as

$$(x \mid y) = E(xy).$$

Since x, y are linear combinations of the y_i's, their inner product can be calculated from the second-order statistics of the y_i's. In particular, if $x = \sum \alpha_i y_i$, $y = \sum \beta_i y_i$, then

$$E(xy) = E\left\{\left(\sum_i \alpha_i y_i\right)\left(\sum_j \beta_j y_j\right)\right\}.$$

The space H is a finite-dimensional Hilbert space with dimension equal to at most m. (If the matrix $G = [E(y_i y_j)]$ is not positive definite, there will be nontrivial linear combinations $\sum \alpha_i y_i$ having zero norm in the space H. These combinations must be considered equivalent to the zero element, thereby reducing the dimension of H.)

If in the Hilbert space H each random variable has expected value equal to zero, then two vectors x, z are orthogonal if they are uncorrelated: $(x \mid z) = E(x)E(z) = 0$.

The concept of a random variable can be generalized in an important direction. An n-dimensional vector-valued random variable x is an ordered collection of n scalar-valued random variables. Notationally, x is written as a column vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

(the components being random variables). For brevity, a vector-valued random variable is referred to simply as a random vector. In applications, the random variables which are components of a random vector may, for example, be associated with the parameters of a single experiment. Or, they may correspond to results of repeated trials of the same experiment or be elements in a random time series such as barometric pressures on five consecutive days.

A Hilbert space of random vectors can be generated from a given set of random vectors in a manner analogous to that for random variables. Suppose $\{y_1, y_2, \ldots, y_m\}$ is a collection of n-dimensional random vectors. Each element y_i has n components y_{ij}, $j = 1, 2, \ldots, n$, each of which is a random variable with finite variance. We define the Hilbert space $\mathscr{H}$ of n-dimensional random vectors as consisting of all vectors whose components are linear combinations of the components of the y_i's. Thus an arbitrary element y in this space can be expressed as

$$y = K_1 y_1 + K_2 y_2 + \cdots + K_m y_m$$

where the K_i's are real $n \times n$ matrices. The resulting space is, of course, generally larger than that which would be obtained by simply considering linear combinations of the y_i's. This specific compound construction of $\mathscr{H}$, although rarely referred to explicitly in our development, is implicitly responsible for the simplicity of many standard results of estimation theory.

If x and z are elements of $\mathscr{H}$, we define their inner product as

$$(x \mid z) = E\left(\sum_{i=1}^{n} x_i z_i\right),$$

which is the expected value of the n-dimensional inner product. A convenient notation is $(x \mid z) = E(x'z)$.

The norm of an element x in the space of n-dimensional random vectors can be written

$$\| x \| = \{\text{Trace } E(xx')\}^{1/2},$$

where

$$E(xx') = \begin{bmatrix} E(x_1 x_1) & E(x_1 x_2) & \cdots & E(x_1 x_n) \\ E(x_2 x_1) & E(x_2 x_2) & \cdots & E(x_2 x_n) \\ \vdots & & & \vdots \\ E(x_n x_1) & E(x_n x_2) & \cdots & E(x_n x_n) \end{bmatrix}$$

is the expected value of the random matrix dyad xx'. Similarly, we have

$$(x \mid z) = \text{Trace } E(xz').$$

In the case of zero means the matrix $E(xx')$ is simply the covariance matrix of the random variables x_i which form the components of x. These components can, of course, be regarded as elements of a "smaller" Hilbert space of random variables; the covariance matrix above then is seen to be the Gram matrix corresponding to $\{x_1, x_2, \ldots, x_n\}$.

Corresponding to the general case with nonzero means, we define the corresponding covariance matrix by

$$\operatorname{cov}(x) = E[(x - E(x))(x - E(x))'].$$

4.3 The Least-Squares Estimate

The purposes of this and the next two sections are to formulate and to solve three basic linear estimation problems. The three problems are quite similar, differing only slightly in the choice of optimality criterion and in the underlying statistical assumptions.

Suppose a quantity of data consisting of m real numbers has been collected and arranged as the m components of a vector y. Often the nature of the source of the data leads one to assume that the vector y, rather than consisting of m independent components, is a given linear function of a few unknown parameters. If these parameters are arranged as the components of an n-dimensional vector β (where $n < m$), such a hypothesis amounts to assuming that the vector y is of the form $y = W\beta$. The $m \times n$ matrix W is, by assumption, determined by the particular experiment or physical situation at hand and is assumed to be known. The data vector y is known. The problem is to determine the vector β. Since $n < m$, however, it is generally not possible to determine a vector β exactly satisfying $y = W\beta$. A useful alternative consists of determining the value of β which best approximates a solution in the sense of minimizing the norm $\| y - W\beta \|$. If this norm is taken as the standard Euclidean m-space norm, this approach leads to a simple least-squares estimate.

As formulated above, this problem is not a statistical one. It simply amounts to approximating y by a vector lying in the subspace spanned by the n column vectors of the matrix W. However, the technique is often used in a statistical setting and it provides a useful comparison for other statistical methods. For example, the procedure might be applied to finding a straight-line approximation to the data in Figure 4.1 (which may represent experimental data gathered on distance traveled versus time for a body under constant acceleration). We hypothesize a model of the form $s = t\beta$ and choose β by least squares. The y vector in this case is made up of the measured s values; the W matrix consists of a single column made up of the

corresponding t values. The error is due to the unaccounted nonlinearity and the measurement errors. If the hypothesized model were originally of the form $s = t\beta_1 + t^2\beta_2$, a better fit could be made.

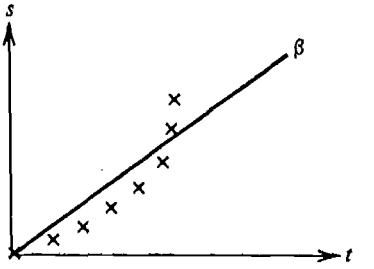

Figure 4.1 Straight-line approximation to data

Theorem 1. (*Least-Squares Estimate*) *Suppose y is an m vector and W an $m \times n$ matrix with linearly independent columns. Then there is a unique n vector $\hat{\beta}$ which minimizes $\|y - W\beta\|$ over all β (the norm taken as the Euclidean m-space norm). Furthermore,*

$$\hat{\beta} = (W'W)^{-1}W'y.$$

Proof. As pointed out above, this problem amounts to approximating y by a linear combination of columns of W. The existence and uniqueness follow immediately from the projection theorem and the independence of the columns of W. Furthermore, the Gram matrix corresponding to the column vectors of W is easily seen to be $W'W$. The vector $W'y$ has as its components inner products of the columns of W with the vector y. Hence the normal equations become

$$W'W\hat{\beta} = W'y.$$

Since the columns of W are assumed to be linearly independent, the Gram matrix $W'W$ is nonsingular and the result follows. ∎

There is an extensive theory dealing with the case where $W'W$ is singular. See the references at the end of this chapter and the discussion of pseudo-inverses in Section 6.11.

Although Theorem 1 is actually only a simple finite-dimensional version of the general approximation problem treated in Chapter 3, the solution is stated here explicitly in matrix notation so that the result can be easily compared with other estimation techniques derived in this chapter.

4.4 Minimum-Variance Unbiased Estimate (Gauss-Markov Estimate)

Assume now that we have arranged an experiment that leads to an m-dimensional data vector y of the form

$$y = W\beta + \varepsilon.$$

In this model W is a known matrix, β is an n-dimensional unknown vector of parameters, and ε is an m-dimensional random vector with zero mean and with covariance $E(\varepsilon\varepsilon') = Q$ which we assume to be positive definite. The vector y can be interpreted as representing the outcome of m inexact measurements, the random vector ε representing the measurement errors. For example, repeated measurements of a single quantity β might be represented as $y_i = \beta + \varepsilon_i$, in which case W would contain a single column with each component equal to unity.

In this section we consider a method for estimating the unknown vector of parameters β from the vector y. In particular we seek a linear estimate β of the form

$$\hat{\beta} = Ky,$$

where K is a constant $n \times m$ matrix.

Since y is the sum of random vector ε and the constant vector $W\beta$, it is itself a random vector. Therefore, both the estimate $\hat{\beta}$ and the error $\hat{\beta} - \beta$ are random vectors with statistics determined by those of ε and the choice of K. A natural criterion for the optimality of the estimation scheme is minimization of the norm of the error, expressed in this case as

$$E[\|\hat{\beta} - \beta\|^2],$$

where $\|\ \|$ is the ordinary Euclidean n-space norm. If, however, this error is written explicitly in terms of the problem variables, we obtain

$$\begin{aligned} (1) \qquad E[\|\hat{\beta} - \beta\|^2] &= E[\|Ky - \beta\|^2] = E[\|KW\beta + K\varepsilon - \beta\|^2] \\ &= \|KW\beta - \beta\|^2 + \text{Trace}\,(KQK'). \end{aligned}$$

The matrix K minimizing this expression is obviously a function of the (unknown) parameter vector β. Therefore, a practical estimation scheme for the parameters β cannot be derived as a solution to the proposed problem.

We observe, however, that if $KW = I$ (the identity matrix), the norm of the error is independent of β. This observation suggests the alternative problem: find the estimate $\hat{\beta} = Ky$ minimizing $E[\|\hat{\beta} - \beta\|^2]$ while satisfying

$$(2) \qquad KW = I.$$

The solution to this problem is independent of β.

The additional requirement $KW = I$ may at first seem to be highly arbitrary and perhaps inappropriate. The requirement has a simple interpretation, however, that tends to defend its introduction. The expected value of the estimate is $E(\hat{\beta}) = E(Ky) = E(KW\beta + K\varepsilon) = KW\beta$. If $KW = I$, the expression reduces to $E(\hat{\beta}) = \beta$. Thus the expected value of the estimate of β is itself β. In general, estimators with this property are said to be *unbiased.* Thus the requirement $KW = I$ leads to an unbiased estimator. Conversely, if we require that an estimator of the form $\hat{\beta} = Ky$ be unbiased in the sense that $E(\hat{\beta}) = \beta$ for all β, it follows that $KW = I$. We seek the unbiased linear estimate of β which minimizes $E(\|\beta - \hat{\beta}\|^2)$.

Before plunging into the detailed analysis of this problem, it behooves us to make some elementary observations concerning its structure. The problem can be written out in terms of the components of $\hat{\beta}$ as that of finding $\hat{\beta}$ to minimize

$$E[\|\hat{\beta} - \beta\|^2] = \sum_{i=1}^{n} E[(\hat{\beta}_i - \beta_i)^2]$$

subject to

$$E(\hat{\beta}_i) = \beta_i, \qquad i = 1, 2, \ldots, n,$$

and

$$\hat{\beta}_i = k_i' y \qquad i = 1, 2, \ldots, n,$$

where k_i' is the i-th row of the matrix K. Therefore, the problem is really n separate problems, one for each $\hat{\beta}_i$. Minimizing each $E(\hat{\beta}_i - \beta_i)^2$ minimizes the sum. Each subproblem can be considered as a constrained minimum norm problem in a Hilbert space of random variables.

An alternative viewpoint is to consider the equivalent deterministic problem of selecting the optimal matrix K. Returning to equations (1) and (2), the problem is one of selecting the $n \times m$ matrix K to

$$\begin{array}{ll} \text{minimize} & \text{Trace } \{KQK'\} \\ \text{subject to} & KW = I. \end{array}$$

This problem can be regarded as a minimum norm problem in the space of matrices (see Problem 3, Chapter 3), or it may be decomposed in a manner analogous to that described for the components of $\hat{\beta}$. The i-th subproblem is

$$\begin{array}{ll} \text{minimize} & k_i' Q k_i \\ \text{subject to} & k_i' w_j = \delta_{ij}, \qquad j = 1, 2, \ldots, n, \end{array}$$

where w_j is the j-th column of W and δ_{ij} is the Kronecker delta function.

Introducing the inner product $(x \mid y)_Q = x'Qy$, the problem becomes

$$\begin{array}{ll} \text{minimize} & (k_i \mid k_i)_Q \\ \text{subject to} & (k_i \mid Q^{-1}w_j)_Q = \delta_{ij}, \qquad j = 1, 2, \ldots, n. \end{array}$$

This is now in the form of the standard minimum norm problem treated in Section 3.10. It is a straightforward exercise to show that

$$k_i = Q^{-1}W(W'Q^{-1}W)^{-1}e_i$$

where e_i is the n-vector with i-th component unity and the rest zero. Note that $W'QW$ is the Gram matrix of the columns of W with respect to $(\,|\,)_Q$. Finally, combining all m of the subproblems, we obtain

$$K' = Q^{-1}W(W'Q^{-1}W)^{-1}.$$

We summarize the above analysis by the following classical theorem.

Theorem 1. *(Gauss-Markov) Suppose* $y = W\beta + \varepsilon$ *where*

$$E(\varepsilon) = 0$$

$$E(\varepsilon\varepsilon') = Q$$

with Q *positive definite. The linear minimum-variance unbiased estimate of* β *is*

$$\hat{\beta} = (W'Q^{-1}W)^{-1}W'Q^{-1}y$$

with corresponding error covariance

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = (W'Q^{-1}W)^{-1}.$$

Proof. The derivation of the estimate is given above. It only remains to calculate the corresponding error covariance

$$\begin{aligned} E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] &= E[(Ky - \beta)(Ky - \beta)'] \\ &= E[K\varepsilon\varepsilon'K'] \\ &= KQK' = (W'Q^{-1}W)^{-1}W'Q^{-1}QQ^{-1}W(W'Q^{-1}W)^{-1} \\ &= (W'Q^{-1}W)^{-1}. \blacksquare \end{aligned}$$

As an aside, it might be mentioned that the justification for the terminology "minimum variance unbiased" rather than "minimum covariance trace unbiased" is that for each i, $\hat{\beta}_i$ is the minimum-variance unbiased estimate of β_i. In other words, the Gauss-Markov estimate provides a minimum-variance unbiased estimate of each component rather than merely a vector optimal in the sense of minimizing the sum of the individual variances.

A striking property of the result of the Gauss-Markov theorem is that if $E(\varepsilon\varepsilon') = I$, the linear, minimum-variance unbiased estimate is identical with the least-squares estimate of Section 4.3. It is clear that the least-squares technique and the Gauss-Markov technique are intimately related. However, a fundamental difference between the two approaches is that least squares is a single elementary minimum norm problem while the Gauss-Markov problem is actually n separate minimum norm problems.

4.5 Minimum-Variance Estimate

In the preceding two sections, the vector β was assumed to be a vector of unknown parameters. Presumably these parameters could have assumed any value from $-\infty$ to $+\infty$; we, as experimenters, had absolutely no prior information concerning their values. In many situations, however, we do have prior information and it is then perhaps more meaningful to regard the unknown vector β as a random variable with known mean and covariance. In such situations, this *a priori* statistical information can be exploited to produce an estimate of lower error variance than the minimum-variance unbiased estimate.

In view of this observation, in this section we again consider estimation of β from measurements of the form

$$y = W\beta + \varepsilon$$

but in this case both β and ε are random vectors. The criterion for optimality is simply the minimization of $E[\|\hat{\beta} - \beta\|^2]$.

We begin by establishing an important theorem which applies in a somewhat more general setting than that described above but which is really only an application of the normal equations.

Theorem 1. (*Minimum-Variance Estimate*) *Let y and β be random vectors (not necessarily of the same dimension). Assume $[E(yy')]^{-1}$ exists. The linear estimate $\hat{\beta}$ of β, based on y, minimizing $E[\|\hat{\beta} - \beta\|^2]$ is*

$$\hat{\beta} = E(\beta y')[E(yy')]^{-1}y,$$

with corresponding error covariance matrix

$$\begin{aligned} E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] &= E(\beta\beta') - E(\hat{\beta}\hat{\beta}') \\ &= E(\beta\beta') - E(\beta y')[E(yy')]^{-1}E(y\beta'). \end{aligned}$$

Proof. This problem, like that in the last section, decomposes into a separate problem for each component β_i. Since there are no constraints, the i-th subproblem is simply that of finding the best approximation of β_i within the subspace generated by the y_j's.

Writing the optimal estimate as $\hat{\beta} = Ky$, where K is an $n \times m$ matrix, the i-th subproblem is equivalent to the problem of selecting the i-th row of K which, in turn, gives the optimal linear combination of the y_j's for $\hat{\beta}_i$. Hence, each row of K satisfies the normal equations corresponding to projecting β_i into the y_j's. The normal equations for all n subproblems can be written simultaneously in the matrix form

$$[E(yy')]K' = E(y\beta')$$

from which it follows that

$$K = E(\beta y')[E(yy')]^{-1},$$

which is the desired result. Proof of the formula for the error covariance is obtained by direct substitution. ∎

Note that, as in the previous section, the term minimum variance applied to these estimates can be taken to imply that each component of β is estimated by a minimum-variance estimator. Also note that if both β and y have zero means, the estimate is unbiased in the sense that $E(\hat{\beta}) = E(\beta) = \theta$. If the means are not zero, we usually broaden the class of estimators to include the form $\hat{\beta} = Ky + b$ where b is an appropriate constant vector. This matter is considered in Problem 6.

Returning to our original purpose, we now apply the above theorem to a revised form of our standard problem.

Corollary 1. *Suppose*

$$y = W\beta + \varepsilon$$

where y is a known m-dimensional data vector, β is an n-dimensional unknown random vector, ε is an unknown m-dimensional random error vector, W is a known $m \times n$ constant matrix, and

$$E(\varepsilon\varepsilon') = Q$$

$$E(\beta\beta') = R, \qquad E(\varepsilon\beta') = 0.$$

We assume that R and Q are positive-semidefinite matrices (of appropriate size) and that $WRW' + Q$ is nonsingular. Then the linear estimate $\hat{\beta}$ of β minimizing $E[\|\hat{\beta} - \beta\|^2]$ is

$$\hat{\beta} = RW'(WRW' + Q)^{-1}y$$

with error covariance

$$E[(\beta - \hat{\beta})(\beta - \hat{\beta})'] = R - RW'(WRW' + Q)^{-1}WR.$$

Proof. It is easily computed that $E(yy') = WRW' + Q$ and that $E(\beta y') = RW'$ from which the result follows by the minimum-variance theorem. ∎

A significant difference between the estimation problem treated above and those of previous sections is that the number of observations, m, need not be at least as large as the number of unknowns, n. The estimate in Corollary 1 exists if the $m \times m$ matrix $(WRW' + Q)^{-1}$ exists; this is the case for any m if, for instance, Q is positive definite. Physically the reason for this property is that since β is now a random variable, it is possible to form a meaningful estimate (namely zero) even with no measurements. Every new measurement simply provides additional information which may modify our original estimate.

Another unique feature of this estimate is that for $m < n$ there need be no measurement error. Thus, we may have $Q = 0$, because as long as WRW' is positive definite the estimate still exists.

Example 1. Consider β to be a random two-dimensional vector with mean zero and covariance I (the identity matrix). Such a vector can be thought of as having an expected length of unity but with a completely random angle as measured from any given axis. Suppose that a single perfect measurement y of the first component β_1 of β has been obtained and from this we seek the best estimate of β. The situation is illustrated in Figure 4.2.

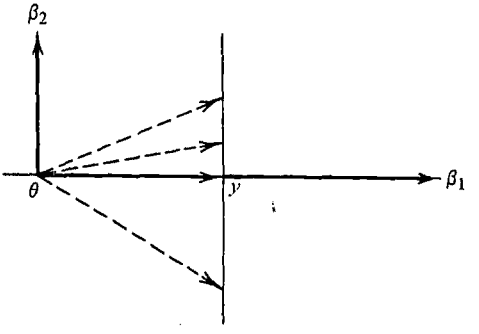

Figure 4.2 Estimation from one component

Intuitively, we choose our estimate somewhere on the vertical line at $\beta_1 = y$. Having no reason to favor one angle over another, we take the shortest vector since it is closest to the mean. Hence we select the horizontal vector that meets the vertical line. This solution can be verified by application of Corollary 1.

By some matrix manipulations, the result of Corollary 1 can be translated into a form that more closely resembles our earlier results.

Corollary 2. *The estimate given by Corollary 1 can be expressed in the alternative form*

$$\hat{\beta} = (W'Q^{-1}W + R^{-1})^{-1}W'Q^{-1}y$$

with corresponding error covariance

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = (W'Q^{-1}W + R^{-1})^{-1}.$$

Proof. The matrix identity

$$RW'(WRW' + Q)^{-1} = (W'Q^{-1}W + R^{-1})^{-1}W'Q^{-1}$$

is easily established by postmultiplying by $(WRW' + Q)$ and premultiplying by $(W'Q^{-1}W + R^{-1})$. This establishes the equivalence of the formula for the estimate $\hat{\beta}$.

From Corollary 1 we have

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = R - RW'(WRW' + Q)^{-1}WR$$

which becomes, by application of the above matrix identity,

$$\begin{aligned} E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] &= R - (W'Q^{-1}W + R^{-1})^{-1}W'Q^{-1}WR \\ &= (W'Q^{-1}W + R^{-1})^{-1}\{(W'Q^{-1}W + R^{-1}) \times R \\ &\quad - W'Q^{-1}WR\} \\ &= (W'Q^{-1}W + R^{-1})^{-1}. \blacksquare \end{aligned}$$

In this form the minimum-variance estimate can be easily compared with the least-squares and Gauss-Markov estimates. In particular, note that if $R^{-1} = 0$ (corresponding to infinite variance of the *a priori* information concerning β), the result of Corollary 2 is identical with the Gauss-Markov estimate. Thus the Gauss-Markov estimate is simply a limiting case of the minimum-variance estimate and, hence, further study of minimum-variance estimates, for the most part, also applies to Gauss-Markov estimates.

4.6 Additional Properties of Minimum-Variance Estimates

In this section we investigate minimum-variance estimation in more detail by considering three problems: the first is to deduce the optimal estimate of a linear function of β, the second is to determine how the optimal estimate of β changes if the optimality criterion is a more general quadratic form, and the third is to determine how an estimate of β is changed if additional measurement data become available. We shall see that these problems are related and that all three have strikingly simple solutions.

Theorem 1. *The minimum-variance linear estimate of a linear function of β, based on the random vector y, is equal to the same linear function of the minimum-variance linear estimate of β; i.e., given an arbitrary $p \times n$ matrix T, the best estimate of $T\beta$ is $T\hat{\beta}$.*

Proof. The result can be obtained from the observation, made in the last section, that the linear estimate $\hat{\beta}$ minimizing $E[\|\beta - \hat{\beta}\|^2]$ actually minimizes $E(\beta_i - \hat{\beta}_i)^2$ for each component β_i of β. We leave it to the reader to complete a proof along these lines.

An alternate proof can be obtained directly from the projection theorem by deriving the optimal estimate of $T\beta$ and comparing the result with $T\hat{\beta}$. If Γy is the optimal estimate of $T\beta$, we must have

$$E[y(T\beta - \Gamma y)'] = 0$$

and, hence, in matrix form the normal equations for the columns of Γ are

$$[E(yy')]\Gamma' = E[y\beta'T']$$

so that

$$\Gamma = TE(\beta y')[E(yy')]^{-1}$$

which, by comparison with the minimum-variance estimate of β, yields the desired result. ∎

Another property of linear minimum-variance estimates, which is closely related to the property considered above and which again can be regarded as a simple consequence of the componentwise optimality of the estimate, is that the estimate is optimal for any positive-semidefinite quadratic optimality criterion.

Theorem 2. *If $\hat{\beta} = Ky$ is the linear minimum-variance estimate of β, then $\hat{\beta}$ is also the linear estimate minimizing $E[(\beta - \hat{\beta})'P(\beta - \hat{\beta})]$ for any positive-semidefinite $n \times n$ matrix P.*

Proof. Let $P^{1/2}$ be the unique positive-semidefinite square root of P. According to Theorem 1, $P^{1/2}\hat{\beta}$ is the minimum-variance estimate of $P^{1/2}\beta$ and, hence, $\hat{\beta}$ minimizes

$$E[\|P^{1/2}\hat{\beta} - P^{1/2}\beta\|^2] = E[(\hat{\beta} - \beta)'P(\hat{\beta} - \beta)]. \quad ∎$$

Finally, we consider the problem of updating an optimal estimate of β if additional data become available. This result is of fundamental practical importance in a number of modern sequential estimation procedures such as the recursive estimation of random processes discussed in Section 4.7. Like the two properties discussed above, the answer to the updating problem is extremely simple.

The procedure is based on the simple orthogonality properties of projection in Hilbert space. The idea is illustrated in Figure 4.3. If Y_1 and Y_2 are subspaces of a Hilbert space, the projection of a vector β onto the subspace $Y_1 + Y_2$ is equal to the projection onto Y_1 plus the projection onto $\tilde{Y}_2$ where $\tilde{Y}_2$ is orthogonal to Y_1 and is chosen so that $Y_1 \oplus \tilde{Y}_2 = Y_1 + Y_2$. Furthermore, if Y_2 is generated by a finite set of vectors, the differences between these vectors and their projections onto Y_1 generate $\tilde{Y}_2$.

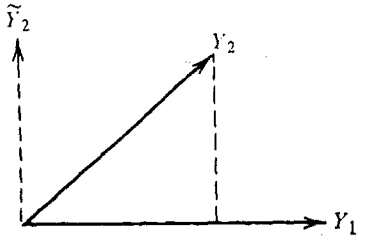

Figure 4.3

For clarity and simplicity, the following theorem is stated in terms of a single unknown random variable β rather than an n-dimensional random vector. Since minimum-variance estimation of an n-dimensional random vector is merely n separate problems, the theorem statement easily carries over to the more general case.

Theorem 3. *Let β be a member of a Hilbert space H of random variables and let $\hat{\beta}_1$ denote its orthogonal projection on a closed subspace Y_1 of H. (Thus, $\hat{\beta}_1$ is the best estimate of β in Y_1.) Let y_2 be an m-vector of random variables generating a subspace Y_2 of H, and let $\hat{y}_2$ denote the m-dimensional vector of the projections of the components of y_2 onto Y_1. (Thus, $\hat{y}_2$ is the vector of best estimates of y_2 in Y_1.) Let $\tilde{y}_2 = y_2 - \hat{y}_2$.*

Then the projection of β onto the subspace $Y_1 + Y_2$, denoted $\hat{\beta}$, is

$$\hat{\beta} = \hat{\beta}_1 + E(\beta \tilde{y}_2')[E(\tilde{y}_2 \tilde{y}_2')]^{-1} \tilde{y}_2 .$$

In other words, $\hat{\beta}$ is $\hat{\beta}_1$ plus the best estimate of β in the subspace $\tilde{Y}_2$ generated by $\tilde{y}_2$.

Proof. It is clear that $Y_1 + Y_2 = Y_1 \oplus \tilde{Y}_2$ and that $\tilde{Y}_2$ is orthogonal to Y_1. The result then follows immediately since the projection onto the sum of subspaces is equal to the sum of the individual projections if the subspaces are orthogonal. ∎

An intuitive interpretation of this result is that if we have an estimate $\hat{\beta}_1$ based on measurements generating Y_1, then when receiving another set of

measurements we should substract out from these measurements that part that could be anticipated from the result of the first measurements. In other words, the updating must be based on that part of the new data which is orthogonal to the old data.

Example 1. Suppose that an optimal estimate $\hat{\beta}$ of a random n-vector β has been formed on the basis of past measurements and that

$$E[(\beta - \hat{\beta})(\beta - \hat{\beta})'] = R.$$

Given additional measurements of the form

$$y = W\beta + \varepsilon$$

where ε is a random vector which has zero mean and which is uncorrelated with both β and the past measurements, we seek the updated optimal estimate $\hat{\hat{\beta}}$ of β.

The best estimate of y based on the past measurements is $\hat{y} = W\hat{\beta}$ and thus $\tilde{y} = y - W\hat{\beta}$. Hence, by Theorem 3,

$$\hat{\hat{\beta}} = \hat{\beta} + E(\beta\tilde{y}')[E(\tilde{y}\tilde{y}')]^{-1}\tilde{y}$$

which works out to be

$$\hat{\hat{\beta}} = \hat{\beta} + RW'[WRW' + Q]^{-1}(y - W\hat{\beta}).$$

The error covariance is

$$E[(\beta - \hat{\hat{\beta}})(\beta - \hat{\hat{\beta}})'] = R - RW'[WRW' + Q]^{-1}WR.$$

4.7 Recursive Estimation

In many applications we are led, quite naturally, to consider a sequence of random variables that occur consecutively in time. For example, daily temperature measurements, the Dow-Jones stock averages, or a sequence of measurements of the position of a maneuvering aircraft can be regarded as sequences of random variables. We define a *discrete random process* as any sequence of random variables.

There are a number of important estimation problems associated with random processes including: prediction, which is estimation of future values of the process from past observations; filtering, which is estimation of the present value of a random process from inexact measurements of the process up to the present time; or, more generally, estimation of one random process from observations of a different but related process. If we require linear minimum-variance estimates, these estimation problems are only special cases of the theory developed earlier in this chapter.

It is customary to depart from our previous notation slightly by indexing the sequence of random variables that compose a discrete process by the

notation $x(k)$ rather than by subscripts. Unless explicitly stated otherwise, all random variables in this section are assumed to be real and to have zero means.

The starting point for the recursive approach to estimation is a representation of random processes that explicitly characterizes the manner in which the process is generated in time. We begin by describing this representation.

Definition. A random process $\{x(k)\}$ is said to be *orthogonal* or *white* if

$$E[x(j)x(k)] = \alpha_j \cdot \delta_{jk}.$$

The process is *orthonormal* if, in addition, $\alpha_j = 1$.

We assume that underlying every observed random process is an orthogonal process in the sense that the variables of the observed process are linear combinations of past values of the orthogonal process. In other words, the given observed process results from the operation on an orthogonal process by a linear-processing device acting in real time.

Example 1. (Moving Average Process) Let $\{u(k)\}_{k=-\infty}^{\infty}$ be an orthornormal random process, and suppose that $\{x(k)\}_{k=-\infty}^{\infty}$ is generated by the formula

$$x(k) = \sum_{j=1}^{\infty} a_j u(k-j)$$

where the constants a_j satisfy $\sum_{j=1}^{\infty} |a_j|^2 < \infty$. The process can be regarded as a moving average of past values of the underlying orthonormal process.

Example 2. (Autoregressive Scheme of Order 1) Let $\{u(k)\}_{k=-\infty}^{\infty}$ be an orthonormal process and suppose that $\{x(k)\}$ is generated by the formula

$$x(k) = ax(k-1) + u(k-1), \qquad |a| < 1.$$

This process, defined by a first-order difference equation, is equivalent to the moving average process

$$x(k) = \sum_{j=1}^{\infty} a^{j-1} u(k-j).$$

Example 3. (Finite-Difference Scheme or Autoregressive Scheme of Order n) As a generalization of the previous example, we imagine that $\{x(k)\}$ is generated from the orthonormal process $\{u(k)\}$ by the finite-difference formula

$$x(k) + a_1 x(k-1) + \cdots + a_n x(k-n) = b_1 u(k-1) + \cdots + b_n u(k-n).$$

In order that the formula represents a stable system (so that $E[x^2(k)]$ remains finite when the formula is assumed to have been operating over the

infinite past), we require that the characteristic polynomial equation

$$s^n + a_1 s^{n-1} + \cdots + a_n = 0$$

has its roots within the unit circle in the complex plane. Alternatively, we may assume that n initial random variables, say $x(0), x(-1), \ldots, x(-n+1)$, have been specified and regard the difference formula as generating $[x(k)]$ for positive k only. In that case, stability is immaterial, although in general $E[x^2(k)]$ may grow without bound as $k \to \infty$. By hypothesizing time-varying coefficients in a finite-difference scheme such as this, we can generate a large class of random processes.

There is a great deal of physical motivation behind these models for random processes. It is believed that basic randomness at the microscopic scale including electron emissions, molecular gas velocities, and elementary particle fission are basically uncorrelated processes. When their effects are observed at the macroscopic scale with, for example, a voltmeter, we obtain some average of the past microscopic effects.

The recursive approach to estimation of random processes is based on a model for the process similar to that in Example 3. However, for convenience and generality, we choose to represent the random process as being generated by a first-order vector difference equation rather than an n-th order scalar difference equation. This model accommodates a larger number of practical situations than the scalar model and simplifies the notation of the analysis.

Definition. An *n-dimensional dynamic model of a random process* consists of the following three parts:

1. A vector difference equation

$$x(k+1) = \Phi(k)x(k) + u(k), \qquad k = 0, 1, 2, \ldots,$$

where $x(k)$ is an n-dimensional state vector, each component of which is a random variable, $\Phi(k)$ is a known $n \times n$ matrix, and $u(k)$ is an n-dimensional random vector input of mean zero satisfying $E[u(k)u'(l)] = Q(k)\delta_{kl}$.

2. An initial random vector $x(0)$ together with an initial estimate $\hat{x}(0)$ having covariance $E[(\hat{x}(0) - x(0))(\hat{x}(0) - x(0))'] = P(0)$.

3. Measurements of the process, assumed to be of the form

$$v(k) = M(k)x(k) + w(k), \qquad k = 0, 1, 2, \ldots,$$

where $M(k)$ is an $m \times n$ matrix and $w(k)$ is an m-dimensional random measurement error having mean zero and satisfying

$$E[w(k)w'(j)] = R(k)\delta_{kj}$$

where $R(k)$ is positive definite.

In addition, it is assumed that the random vectors $x(0)$, $u(j)$ and $w(k)$ are all uncorrelated for $j \geq 0, k \geq 0$.

This model covers most of the examples discussed previously as well as several other important situations. In many applications, the matrices $\Phi(k)$, $Q(k)$, $M(k)$, and $R(k)$ are independent of k.

The estimation problem is that of obtaining the linear minimum-variance estimate of the state vector x from the measurements v. We introduce the notation[1] $\hat{x}(k \mid j)$ for the optimal estimate of $x(k)$, given the observations v up to instant j. Thus $\hat{x}(k \mid j)$ is the appropriate projection of $x(k)$ onto the space $V(j)$ generated by the random m-vectors $v(0), v(1), \ldots, v(j)$.

We are concerned exclusively with the case $k \geq j$—the case corresponding to either prediction of some future value of the state or estimation of the present state. Estimation of past values of the state vector is referred to as the smoothing problem; although in principle it is not substantially different than prediction, the detailed calculations are a good deal more messy. Solution to the estimation problem is quite straightforward, requiring primarily the successive application of the updating procedure of Section 4.6.

Theorem 1. (*Solution to Recursive Estimation Problem—Kalman*) *The optimal estimate $\hat{x}(k+1 \mid k)$ of the random state vector may be generated recursively according to the equation*

$$\hat{x}(k+1 \mid k) = \Phi(k)P(k)M'(k)[M(k)P(k)M'(k) + R(k)]^{-1} \times [v(k) - M(k)\hat{x}(k \mid k-1)] + \Phi(k)\hat{x}(k \mid k-1) \tag{1}$$

where the $n \times n$ matrix $P(k)$ is the covariance of $\hat{x}(k \mid k-1)$ which itself is generated recursively by

$$P(k+1) = \Phi(k)P(k)\{I - M'(k)[M(k)P(k)M'(k) + R(k)]^{-1}M(k)P(k)\}\Phi'(k) + Q(k). \tag{2}$$

The initial conditions for these recurrence equations are the initial estimate $\hat{x}(0 \mid -1) = \hat{x}(0)$ and its associated covariance $P(0)$.

Proof. Suppose that $v(0), v(1), \ldots, v(k-1)$ have been measured and that the estimate $\hat{x}(k \mid k-1)$ together with the covariance matrix $P(k) = E[(\hat{x}(k \mid k-1) - x(k))(\hat{x}(k \mid k-1) - x(k))']$ have been computed. In other words, we have the projection of $x(k)$ onto the subspace $V(k-1)$. At time k, we obtain the measurements

$$v(k) = M(k)x(k) + w(k)$$

which gives us additional information about the random vector $x(k)$. This

[1] In this section this notation should not be confused with that of an inner product.

is exactly the situation considered in Example 1, Section 4.6. The updated estimate of $x(k)$ is

$$\hat{x}(k \mid k) = \hat{x}(k \mid k-1) + P(k)M'(k)[M(k)P(k)M'(k) + R(k)]^{-1} \times [v(k) - M(k)\hat{x}(k \mid k-1)] \tag{3}$$

with associated error covariance

$$P(k \mid k) = P(k) - P(k)M'(k)[M(k)P(k)M'(k) + R(k)]^{-1}M(k)P(k). \tag{4}$$

Based on this optimal estimate of $x(k)$, we may now compute the optimal estimate $\hat{x}(k+1 \mid k)$ of $x(k+1) = \Phi(k)\hat{x}(k) + u(k)$, given the observation $v(k)$. We do this by noting that by Theorem 1, Section 4.6, the optimal estimate of $\Phi(k)x(k)$ is $\Phi(k)\hat{x}(k \mid k)$, and since $u(k)$ is orthogonal (uncorrelated) to $v(k)$ and $x(k)$, the optimal estimate of $x(k+1)$ is

$$\hat{x}(k+1 \mid k) = \Phi(k)\hat{x}(k \mid k). \tag{5}$$

The error covariance of this estimate is

$$P(k+1) = \Phi(k)P(k \mid k)\Phi'(k) + Q(k). \tag{6}$$

Substitution of equation (3) into (5) and of (4) into (6) leads directly to (1) and (2). ∎

Equations (1) and (2) may at first sight seem to be quite complicated, but it should be easily recognized that they merely represent the standard minimum-variance formulae together with a slight modification due to the updating process. Furthermore, although these equations do not allow for simple hand computations or analytic expressions, their recursive structure is ideally suited to machine calculation. The matrices $P(k)$ can be precomputed from equation (2); then, when filtering, only a few calculations must be made as each new measurement is received.

Theorem 1 treats only estimates of the special form $\hat{x}(k+1 \mid k)$ rather than $\hat{x}(j \mid k)$ for arbitrary j. Solutions to the more general problem, however, are based on the estimate $\hat{x}(k+1 \mid k)$. See Problems 15 and 16.

4.8 Problems

1. A single scalar β is measured by m different people. Assuming uncorrelated measurement errors of equal variance, show that the minimum-variance unbiased estimate of β is equal to the average of the m measurement values.
2. Three observers, located on a single straight line, measure the angle of their line-of-sight to a certain object (see Figure 4.4). It is desired to estimate the true position of the object from these measurements.

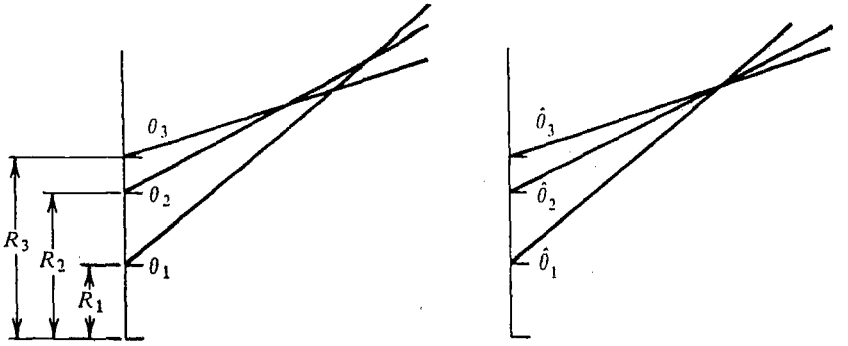

Figure 4.4 A triangulation problem

Assuming that the angles measured, θ_i, $i = 1, 2, 3$, are sufficiently accurate so that "small angle approximations" apply to the deviations, show that the estimated position defined by $\hat{\theta}_i$, $i = 1, 2, 3$, which minimizes $\sum_{i=1}^{3} (\theta_i - \hat{\theta}_i)^2$ is given by

$$\hat{\theta}_1 = \theta_1 + k(R_2 - R_3)S_1^2$$
$$\hat{\theta}_2 = \theta_2 + k(R_3 - R_1)S_2^2$$
$$\hat{\theta}_3 = \theta_3 + k(R_1 - R_2)S_3^2$$

where S_i = secant θ_i and k is chosen so that $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, define a single point of intersection.

3. A certain type of mass-spectrometer produces an output graph similar to that shown in Figure 4.5. Ideally, this curve is made up of a linear combination of several identical but equally displaced pulses. In other words,

$$s(t) = \sum_{i=0}^{n-1} \beta_i a(t - i)$$

where the β_i's are unknown constants and $a(t)$ is a known function. Assuming that $\int_{-\infty}^{\infty} a(t)a(t - i)\,dt = p^i$, show that the least-squares estimate of the β_i's, given an arbitrary measurement curve $s(t)$, is

$$\hat{\beta}_i = \frac{1}{1 - p^2} \int_{-\infty}^{\infty} b_i(t)s(t)\,dt$$

where

$b_i(t) =$

$$\begin{cases} a(t - i) - pa(t - i - 1) & i = 0 \\ (1 + p^2)a(t - i) - pa(t - i - 1) - pa(t - i + 1) & 0 < i < n - 1 \\ a(t - i) - pa(t - i + 1) & i = n - 1 \end{cases}$$

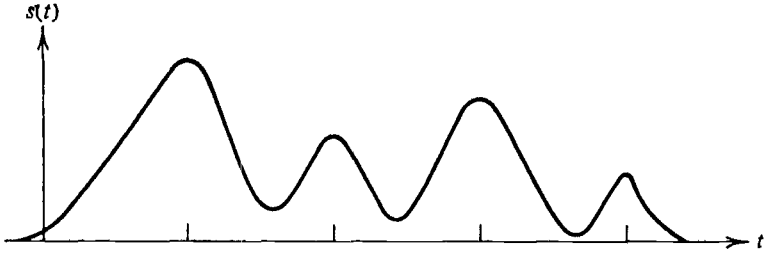

Figure 4.5 Mass-spectrometer data

4. Let β be an n-dimensional random vector of zero mean and positive-definite covariance matrix Q. Suppose measurements of the form $y = W\beta$ are made where the rank of W is m. If $\hat{\beta}$ is the linear minimum variance estimate of β based on y, show that the covariance of the error $\beta - \hat{\beta}$ has rank $n - m$.
5. Assume that the measurement vector y is obtained from β by

$$y = W\beta + \varepsilon$$

where

$$E(\beta\beta') = R, \qquad E(\varepsilon\varepsilon') = Q, \qquad E(\beta\varepsilon') = S.$$

Show that the minimum-variance estimate of β based on y is

$$\hat{\beta} = (RW' + S)(WRW' + WS + S'W' + Q)^{-1}y.$$

6. Let β, y be random vectors with $E(\beta) = \beta_0$, $E(y) = y_0$, and finite-covariance matrices. Show that the minimum-variance estimate of β of form

$$\hat{\beta} = Ky + b,$$

where b is a constant vector, is

$$\hat{\beta} = \beta_0 + E[(\beta - \beta_0)(y - y_0)']\{E[(y - y_0)(y - y_0)']\}^{-1}(y - y_0).$$

7. Let $\hat{\beta} = Ky$ be the minimum-variance linear estimate of a random vector β based on the random vector y. Show that

$$E[(\beta - \hat{\beta})(\beta - \hat{\beta})'] = E[\beta\beta'] - E[\hat{\beta}\hat{\beta}'].$$

8. In this problem we develop the rudiments of the theory of linear regression. Suppose that associated with an experiment there are two random variables y and x. If the outcomes of several measurements of y and x are plotted on a two-dimensional graph, the result may look somewhat like that shown in Figure 4.6. These results could be effectively summarized by saying that y is approximately a linear function of x. So y would be described by the equation $y = a + bx$ which

is represented by the dashed line in the figure. A natural way to choose the appropriate dashed line is to choose that line which minimizes the total sum of the squared errors $\sum_{i=1}^{N} e_i^2$ where $e_i = y_i - (a + bx_i)$ is the vertical distance between an observation point on the graph and the dashed line.

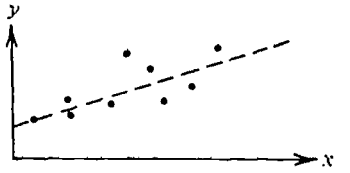

Figure 4.6 Regression

(a) Show that the best linear approximation is given by

$$y \simeq \bar{y} + b(x - \bar{x})$$

where

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \qquad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad b = \frac{\sum_{i=1}^{N} y_i x_i - N\bar{y}\bar{x}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}.$$

(b) Show that b may be alternatively expressed as

$$b = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}.$$

9. With the same terminology as in Problem 8, suppose now that the random variable y can be represented as

$$y = \alpha + \beta x + \varepsilon$$

where ε is a random variable with zero mean, variance σ_ε^2, and independent of x. We are hypothesizing the existence of a linear relation between x and y together with an additive error variable. The values of a and b found in Problem 8 are estimates of the parameters α and β.

Suppose the outcomes of N observations produce $x_1, x_2, \ldots, x_N$. (The x_i's are thus fixed in the discussion that follows.)

(a) Show that

$$b = \beta + \frac{\sum_{i=1}^{N} (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

so that $E(b) = \beta$.

(b) Show that $E(a) = \alpha$.

(c) Show that

$$\text{Var}(b) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}.$$

(d) What is Var (b)?

10. Suppose a large experiment producing data of the form

$$y = W\beta + \varepsilon$$

is planned with y an m-dimensional data vector, β an n-dimensional vector of unknown parameters, and ε a random error vector with $E(\varepsilon) = \theta$, $E(\varepsilon\varepsilon') = I$. Before the experiment the matrix $(W'W)^{-1}$, which is of high order, is calculated. Unfortunately, during the experiment the last component of y is not obtained and the Gauss-Markov estimate of β must be computed on the basis of the remaining components. Show that the inverse need not be recalculated but that $\hat{\beta}$ can be determined using $(W'W)^{-1}$ and the last row of W.

11. Show that, analogous to Theorem 1, Section 4.6, the Gauss-Markov estimate of $T\beta$ is $T\hat{\beta}$ where $\hat{\beta}$ is the Gauss-Markov estimate of β.

12. Show that, analogous to Theorem 2, Section 4.6, the Gauss-Markov estimate $\hat{\beta}$ of β is the linear minimum-variance unbiased estimate minimizing $E[(\beta - \hat{\beta})P(\beta - \hat{\beta})']$ for any positive-semidefinite matrix P.

13. It is the purpose of this problem to show that Gauss-Markov estimates can be updated in a manner analogous to that of minimum-variance estimates. As a by-product, a formal connection between the two kinds of estimation techniques is obtained.

Let y_1 and y_2 be two measurement vectors of an n-dimensional vector of unknown parameters β; say,

$$y_1 = W_1\beta + \varepsilon_1$$

$$y_2 = W_2\beta + \varepsilon_2$$

where $E(\varepsilon_1) = \theta$, $E(\varepsilon_2) = \theta$, $E(\varepsilon_1\varepsilon_1') = Q_1$, $E(\varepsilon_2\varepsilon_2') = Q_2$, and $E(\varepsilon_1\varepsilon_2') = 0$. Assuming that the dimension m_1 of y_1 is at least n and that $W_1'Q_1^{-1}W_1$ is nonsingular, show that the Gauss-Markov estimated of β based on y_1 and y_2 can be obtained by first obtaining the

Gauss-Markov estimate, as well as the corresponding error covariance of β based on y_1, and then updating it with the minimum-variance estimator based on y_2.

14. A satellite is put into orbit about the earth. Its initial state vector (whose components are position and velocity coordinates) is not known precisely and it is desired to estimate it from later measurements of the satellite's position. The equations of motion are:

$$x(k+1) = \Phi x(k)$$

where $x(k)$ is the n-dimensional state vector and Φ is an $n \times n$ matrix. It is assumed that

$$E[x(0)] = \theta, \qquad E[x(0)x'(0)] = P.$$

The measurements are of the form

$$v(k) = Mx(k) + \varepsilon(k)$$

where M is an $m \times n$ matrix ($m < n$) and $E(\varepsilon(k)) = \theta$, $E[\varepsilon(k)\varepsilon'(j)] = Q\delta_{jk}$. Develop the recursive equations for the minimum-variance linear estimate of $x(0)$ given the observations.

15. Given the general dynamical model of a random process as in Section 4.7, show that $\hat{x}(k \mid j) = \Phi(k-1)\Phi(k-2)\ldots\Phi(j+1)\hat{x}(j+1 \mid j)$ for $k > j$.

16. Given the general dynamical model of a random process as in Section 4.7, show how to calculate recursively $\hat{x}(0 \mid j)$, $j > 0$.

REFERENCES

§4.1. For more general approaches to estimation, not restricted to least squares, consult the text by Deutsch [40].

§4.3. The method of least-squares estimation goes back to Legendre and Gauss, the latter having made extensive use of the method for planetary orbit calculations.

§4.6. For a detailed study of minimum-variance unbiased estimates and a general foundation of statistical estimation, see Freeman [56] or Graybill [65]. For additional applications and extensions of the theory, see Chipman [29] and Mann [103].

§4.7. The theory of estimation of random processes was initiated by Wiener [154] and independently by Kolmogoroff who used a Hilbert space approach. This section closely follows Kalman [77] who developed the recursive approach.

§4.8. The mass-spectrometer problem (Problem 3) is treated in Luenberger [98]. For a solution of the smoothing problem (Problem 16), see Rauch, Tung, and Striebel [122].

5

DUAL SPACES

5.1 Introduction

The modern theory of optimization in normed linear space is largely centered about the interrelations between a space and its corresponding dual—the space consisting of all continuous linear functionals on the original space. In this chapter we consider the general construction of dual spaces, give some examples, and develop the most important theorem in this book—the Hahn-Banach theorem.

In the remainder of the book we witness the interplay between a normed space and its dual in a number of distinct situations. Dual space plays a role analogous to the inner product in Hilbert space; by suitable interpretation we can develop results extending the projection theorem solution of minimum norm problems to arbitrary normed linear spaces. Dual space provides the setting for an optimization problem "dual" to a given problem in the original (primal) space in the sense that if one of these problems is a minimization problem, the other is a maximization problem. The two problems are equivalent in the sense that the optimal values of objective functions are equal and solution of either problem leads to solution of the other. Dual space is also essential for the development of the concept of a gradient, which is basic for the variational analysis of optimization problems. And finally, dual spaces provide the setting for Lagrange multipliers, fundamental for a study of constrained optimization problems.

Our approach in this chapter is largely geometric. To make precise mathematical statements, however, it is necessary to translate these geometric concepts into concrete algebraic relations. In this chapter we follow two paths to a final set of algebraic results by considering two different geometrical viewpoints, corresponding to two versions of the Hahn-Banach theorem. The first viewpoint parallels the development of the projection theorem, while the second is based on the idea of separating convex sets with hyperplanes.

LINEAR FUNCTIONALS

5.2 Basic Concepts

First we recall that a functional f on a vector space X is *linear* if for any two vectors $x, y \in X$ and any two scalars α, β there holds $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$.

Example 1. On the space E^n a linear functional can be expressed in the form $f(x) = \sum_{k=1}^{n} \eta_k \xi_k$, for $x = (\xi_1, \xi_2, \ldots, \xi_n)$, where the η_k are fixed scalars. Such a functional is easily seen to be linear. Furthermore, it can be shown that all linear functionals on E^n are of this form.

Example 2. On the space $C[0, 1]$ the functional $f(x) = x(\frac{1}{2})$ is a linear functional.

Example 3. On the space $L_2[0, 1]$ the functional $f(x) = \int_0^1 y(t)x(t)\,dt$, for a fixed $y \in L_2[0, 1]$, is a linear functional.

We are primarily concerned with continuous linear functionals. The reader may verify that the functionals in the above three examples are all continuous.

Proposition 1. *If a linear functional on a normed space X is continuous at a single point, it is continuous throughout X.*

Proof. Assume that f is linear and continuous at $x_0 \in X$. Let $\{x_n\}$ be a sequence from X converging to $x \in X$. Then, by the linearity of f,

$$|f(x_n) - f(x)| = |f(x_n - x + x_0) - f(x_0)|.$$

However, since $x_n - x + x_0 \to x_0$ and since f is continuous at x_0, we have $f(x_n - x + x_0) \to f(x_0)$. Thus, $|f(x_n) - f(x)| \to 0$. ∎

The above result is most often applied to the point θ and continuity thus verified by verifying that $f(x_n) \to 0$ for any sequence tending to θ. Intimately related to the notion of continuity is the notion of boundedness.

Definition. A linear functional f on a normed space is *bounded* if there is a constant M such that $|f(x)| \leq M\|x\|$ for all $x \in X$. The smallest such constant M is called the norm of f and is denoted $\|f\|$. Thus, $\|f\| = \inf\{M : |f(x)| \leq M\|x\|, \text{ all } x \in X\}$.

It is shown below that this definition satisfies the usual requirements of a norm.

Proposition 2. *A linear functional on a normed space is bounded if and only if it is continuous.*

Proof. Suppose first that the linear functional f is bounded. Let M be such that $|f(x)| \leq M\|x\|$ for all $x \in X$. Then if $x_n \to \theta$, we have $|f(x_n)| \leq M\|x_n\| \to 0$. Thus, f is continuous at θ. From Proposition 1 it follows that f is continuous everywhere.

Now assume that f is continuous at θ. Then there is a $\delta > 0$ such that $|f(x)| < 1$ for $\|x\| \leq \delta$. Since for any nonzero $x \in X$, $\delta x/\|x\|$ has norm equal to δ, we have

$$|f(x)| = \left|f\left(\frac{\delta x}{\|x\|}\right)\right| \cdot \frac{\|x\|}{\delta} < \frac{\|x\|}{\delta}$$

and $M = 1/\delta$ serves as a bound for f. ∎

We offer now an example of an unbounded linear functional.

Example 4. On the space of finitely nonzero sequences with norm equal to the maximum of the absolute values of the components, we define, for $x = \{\xi_1, \xi_2, \ldots, \xi_n, 0, 0, \ldots\}$,

$$f(x) = \sum_{k=1}^{n} k\xi_k.$$

The functional f is clearly linear but unbounded.

The linear functionals on a vector space may themselves be regarded as elements of a vector space by introducing definitions of addition and scalar multiplication. Given two linear functionals f_1, f_2 on a space X, we define their sum $f_1 + f_2$ as the functional on X given by $(f_1 + f_2)(x) = f_1(x) + f_2(x)$ for all $x \in X$. Similarly, given a linear functional f, we define αf by $(\alpha f)(x) = \alpha[f(x)]$. The null element in the space of linear functionals is the functional that is identically zero on X. The space of linear functionals defined in this way is called the *algebraic dual* of X. Its definition is independent of any topological structure on X such as might be induced by a norm on X.

Of greater importance for our purposes is the subspace of the algebraic dual consisting of all bounded (i.e., continuous) linear functionals on a normed space X. The space becomes a normed space by assigning the norm according to the last definition.

The norm of a functional f can be expressed in several alternative ways. We have

$$\begin{aligned}\|f\| &= \inf_M \{M : |f(x)| \leq M\|x\|, \text{ for all } x \in X\} \\ &= \sup_{x \neq \theta} \frac{|f(x)|}{\|x\|} \\ &= \sup_{\|x\| \leq 1} |f(x)| \\ &= \sup_{\|x\| = 1} |f(x)|.\end{aligned}$$

The reader should verify these equivalences since they are used throughout this chapter. The norm defined in this way satisfies the usual requirements of a norm: $\|f\| > 0$; $\|f\| = 0$ if and only if $f = \theta$; $\|\alpha f\| = |\alpha|\,\|f\|$; $\|f_1 + f_2\| = \sup_{\|x\| \le 1} |f_1(x) + f_2(x)| \le \sup_{\|x\| \le 1} |f_1(x)| + \sup_{\|x\| \le 1} |f_2(x)| = \|f_1\| + \|f_2\|$.

In view of the preceding discussion, the following definition is justified:

Definition. Let X be a normed linear vector space. The space of all bounded linear functionals on X is called the *normed dual* of X and is denoted X^*. The norm of an element $f \in X^*$ is

$$\|f\| = \sup_{\|x\| \le 1} |f(x)|.$$

Given a normed space X, we usually refer to its normed dual X^* simply as the *dual* of X. As a general rule we denote functionals, linear or not, on a normed space X by f, g, h, etc. However, when in the context of a particular development, certain bounded linear functionals are regarded as elements of the space X^*; they are usually denoted by x_1^*, x_2^*, etc. The value of the linear functional $x^* \in X^*$ at the point $x \in X$ is denoted by $x^*(x)$ or by the more symmetric notation $\langle x, x^* \rangle$ which is introduced in Section 5.6.

Theorem 1. *X^* is a Banach space.*

Proof. Since it has already been established that X^* is a normed linear space, it remains only to show that X^* is complete. For this purpose, let $\{x_n^*\}$ be a Cauchy sequence in X^*. This means that $\|x_n^* - x_m^*\| \to 0$ as $n, m \to \infty$. Now for any $x \in X$, $\{x_n^*(x)\}$ is a Cauchy sequence of scalars since $|x_n^*(x) - x_m^*(x)| \le \|x_n^* - x_m^*\| \cdot \|x\|$. Hence, for each x, there is a scalar $x^*(x)$ such that $x_n^*(x) \to x^*(x)$. The functional x^* defined on all of X in this way is certainly linear since $x^*(\alpha x + \beta y) = \lim x_n^*(\alpha x + \beta y) = \lim [\alpha x_n^*(x) + \beta x_n^*(y)] = \alpha \lim x_n^*(x) + \beta \lim x_n^*(y) = \alpha x^*(x) + \beta x^*(y)$.

Now since $\{x_n^*\}$ is Cauchy, given $\varepsilon > 0$, there is an M such that $|x_n^*(x) - x_m^*(x)| < \varepsilon\|x\|$ all $n, m > M$ and all x; but since $x_n^*(x) \to x^*(x)$, we have $|x^*(x) - x_m^*(x)| < \varepsilon\|x\|$, $m > M$. Thus,

$$\begin{aligned} |x^*(x)| = |x^*(x) - x_m^*(x) + x_m^*(x)| &\le |x^*(x) - x_m^*(x)| + |x_m^*(x)| \\ &\le (\varepsilon + \|x_m^*\|)\|x\| \end{aligned}$$

and x^* is a bounded linear functional. Also from $|x^*(x) - x_m^*(x)| < \varepsilon\|x\|$, $m > M$, there follows $\|x^* - x_m^*\| < \varepsilon$ so that $x_m^* \to x^*$ in X^*. ∎

5.3 Duals of Some Common Banach Spaces

In this section we develop representations of the duals of E^n, l_p, L_p, c_0, and Hilbert space. The dual of $C[a, b]$ is discussed in Section 5.5. These

concrete representations of the duals of various Banach spaces enable us to apply the abstract theory of functional analysis to specific practical problems.

The Dual of E^n. In the space E^n, each vector is an n-tuple of real scalars $x = (\xi_1, \xi_2, \ldots, \xi_n)$ with norm $\|x\| = (\sum_{i=1}^n \xi_i^2)^{1/2}$. Any functional f of the form $f(x) = \sum_{i=1}^n \eta_i \xi_i$ with each η_i a real number is clearly linear. Also, from the Cauchy-Schwarz inequality for finite sequences,

$$|f(x)| = \left|\sum_{i=1}^n \eta_i \xi_i\right| \leq \left(\sum_{i=1}^n \eta_i^2\right)^{1/2}\left(\sum_{i=1}^n \xi_i^2\right)^{1/2} = \left(\sum_{i=1}^n \eta_i^2\right)^{1/2} \|x\|,$$

we see that f is bounded with $\|f\| \leq (\sum_{i=1}^n \eta_i^2)^{1/2}$. However, since for $x = (\eta_1, \eta_2, \ldots, \eta_n)$ equality is achieved in the Cauchy-Schwarz inequality, we must in fact have $\|f\| = (\sum_{i=1}^n \eta_i^2)^{1/2}$.

Now let f be any bounded linear functional on E^n. Define the basis vectors e_i in X by $e_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ with the i-th component 1 and all others 0. Let $\eta_i = x^*(e_i)$. A vector $x = (\xi_1, \xi_2, \ldots, \xi_n)$ may be expressed in terms of the basis as $x = \sum_{i=1}^n \xi_i e_i$. Since f is linear, we have

$$f(x) = \sum \xi_i f(e_i) = \sum \eta_i \xi_i.$$

Thus the dual space X^* of $X = E^n$ is itself E^n in the sense that the space X^* consists of all functionals of the form $f(x) = \sum \eta_i \xi_i$ and the norm on X^* is $\|f\| = (\sum_{i=1}^n \eta_i^2)^{1/2}$.

The Dual of l_p, $1 \leq p < \infty$. The l_p spaces were discussed in Section 2.10. For every p, $1 \leq p < \infty$, we define the conjugate index $q = p/(p-1)$, so that $1/p + 1/q = 1$; if $p = 1$, we take $q = \infty$. We now show that the dual space of l_p is l_q.

Theorem 1. *Every bounded linear functional on l_p, $1 \leq p < \infty$, is representable uniquely in the form*

$$f(x) = \sum_{i=1}^{\infty} \eta_i \xi_i \tag{1}$$

where $y = \{\eta_i\}$ is an element of l_q. Furthermore, every element of l_q defines a member of $(l_p)^*$ in this way, and we have

$$\|f\| = \|y\|_q = \begin{cases} \left(\sum_{i=1}^{\infty} |\eta_i|^q\right)^{1/q} & \text{if } 1 < p < \infty \\ \sup_k |\eta_k| & \text{if } p = 1. \end{cases} \tag{2}$$

Proof. Suppose f is a bounded linear functional on l_p. Define the element $e_i \in l_p$, $i = 1, 2, \ldots$, as the sequence that is identically zero except

for a 1 in the i-th component. Define $\eta_i = f(e_i)$. For any $x = \{\xi_i\} \in l_p$, we have, by the continuity of f, $f(x) = \sum_{i=1}^{\infty} \eta_i \xi_i$.

Suppose first that $1 < p < \infty$. For a given positive integer N define the vector $x_N \in l_p$ having components

$$\xi_i = \begin{cases} |\eta_i|^{q/p} \operatorname{sgn} \eta_i & i \leq N \\ 0 & i > N. \end{cases}$$

Then

$$\|x_N\| = \left(\sum_{i=1}^{N} |\eta_i|^q \right)^{1/p}$$

and

$$f(x_N) = \sum_{i=1}^{N} |\eta_i|^{(q/p)+1} = \sum_{i=1}^{N} |\eta_i|^q.$$

But $|f(x_N)| \leq \|f\| \|x_N\|$; therefore, from the above two expressions, it follows that

$$\left(\sum_{i=1}^{N} |\eta_i|^q \right)^{1/q} \leq \|f\| \qquad \text{for all } N.$$

Hence the sequence $y = \{\eta_i\}$ is an element of l_q, and $\|y\|_q \leq \|f\|$.

Suppose now that $y = \{\eta_i\}$ is an element of l_q. If $x = \{\xi_i\} \in l_p$, then $f(x) = \sum_{i=1}^{\infty} \xi_i \eta_i$ is a bounded linear functional on l_p since, by the Hölder inequality, $|f(x)| \leq \sum_{i=1}^{\infty} |\xi_i \eta_i| \leq \|x\|_p \|y\|_q$, and thus $\|f\| \leq \|y\|_q$. Since $f(e_i) = \eta_i$ in this case, it follows from the previous analysis that $\|y\|_q \leq \|f\|$. Therefore, $\|f\| = \|y\|_q$.

For $p = 1$, $q = \infty$, define x_N by

$$\xi_i = \begin{cases} 0 & i \neq N \\ \operatorname{sgn} \eta_N & i = N. \end{cases}$$

Then $\|x_N\| \leq 1$ and

$$|\eta_N| = f(x_N) \leq \|f\| \, \|x_N\| \leq \|f\|.$$

Thus the sequence $y = \{\eta_i\}$ is bounded by $\|f\|$. Hence, $\|y\|_\infty \leq \|f\|$.

Conversely, if $y = \{\eta_i\} \in l_\infty$, the relation (1) obviously defines an element f of $(l_1)^*$ with $\|f\| \leq \|y\|_\infty$. Since again $f(e_i) = \eta_i$, it follows from above that $\|y\|_\infty \leq \|f\|$ and, hence, $\|f\| = \|y\|_\infty$. ∎

The dual of l_∞ is not l_1.

The Dual of $L_p[0, 1]$, $1 \leq p < \infty$. The L_p spaces were discussed in Section 2.10 and are the function space analogs of the l_p spaces. Arguments similar to those given in Theorem 1 show that for $1 \leq p < \infty$, the dual

space of L_p is L_q; $(1/p + 1/q = 1)$ in the sense that there is a one-to-one correspondence between bounded linear functionals f and elements $y \in L_q$ such that

$$f(x) = \int_0^1 x(t)y(t)\,dt$$

and $\|f\| = \|y\|_q$.

The Dual of c_0. The space c_0 is defined as the space of all infinite sequences $x = \{\xi_i\}$ of real numbers converging to zero. The norm on c_0 is $\|x\| = \max_i |\xi_i|$. Thus, c_0 is a subspace of l_∞.

We leave it to the reader to verify that the dual of c_0 is l_1 in the usual sense with bounded linear functionals represented as

$$f(x) = \sum_{i=1}^{\infty} \xi_i \eta_i, \qquad y = \{\eta_1, \eta_2, \ldots\} \in l_1$$

and $\|f\| = \|y\|$.

The Dual of Hilbert Space. On a Hilbert space the functional $f(x) = (x \mid y)$ for a fixed y is a linear functional in the variable x. The Cauchy-Schwarz inequality $|(x \mid y)| \leq \|x\|\|y\|$ shows that the functional f is bounded with $\|f\| \leq \|y\|$; the relation $f(y) = (y \mid y)$ shows that in fact $\|f\| = \|y\|$. Obviously, distinct vectors y produce distinct functionals f. Thus, in Hilbert space, bounded linear functionals are generated by elements of the space itself. Here we show that all bounded linear functionals on Hilbert space are of this form. The examples for the Hilbert spaces E^n, l_2, and L_2 considered above illustrate this general result.

Theorem 2. (*Riesz-Fréchet*) *If f is a bounded linear functional on a Hilbert space H, there exists a unique vector $y \in H$ such that for all $x \in H$, $f(x) = (x \mid y)$. Furthermore, we have $\|f\| = \|y\|$ and every y determines a unique bounded linear functional in this way.*

Proof. Given a bounded linear functional f, let N be the set of all vectors $n \in H$ for which $f(n) = 0$. The set N is obviously a subspace of H. It is closed since if $n_i \to x$ is a sequence in H with $n_i \in N$, we have $0 = f(n_i) \to f(x)$ by the continuity of f.

If $N = H$, then $f \equiv 0$ and the theorem is proved by taking $y = \theta$.

If $N \neq H$, we may write, according to Theorem 1, Section 3.4, $H = N \oplus N^\perp$, and since $N \neq H$, there is a nonzero vector $z \in N^\perp$. Since z is nonzero and $z \notin N$, necessarily $f(z) \neq 0$. Since $N^\perp$ is a subspace, we may assume that z has been appropriately scaled so that $f(z) = 1$. It will be shown that the vector z is a scalar multiple of the desired vector y.

Given any $x \in H$, we have $x - f(x)z \in N$ since $f[x - f(x)z] = f(x) - f(x)f(z) = 0$. Since $z \perp N$, we have $(x - f(x)z \mid z) = 0$ or $(x \mid z) = f(x)\|z\|^2$ or $f(x) = (x \mid z/\|z\|^2)$. Thus, defining $y = z/\|z\|^2$, we have $f(x) = (x \mid y)$.

The vector y is clearly unique since if y' is any vector for which $f(x) = (x \mid y')$ for all x we have $(x \mid y) = f(x) = (x \mid y')$, or $(x \mid y - y') = 0$ for all x which according to Lemma 2, Section 3.2 implies $y' = y$.

It was shown in the discussion preceding the theorem that $\|f\| = \|y\|$. ∎

EXTENSION FORM OF THE HAHN-BANACH THEOREM

5.4 Extension of Linear Functionals

The Hahn-Banach theorem, the most important theorem for the study of optimization in linear spaces, can, like so many important mathematical results, be stated in several equivalent ways each having its own particular conceptual advantage. The two classical versions of the theorem, referred to as the "extension form" and the "geometric form," play a fundamental role in the theory of this book. The extension form proved in this section serves as an appropriate generalization of the projection theorem from Hilbert space to normed space and thus provides a means for generalizing many of our earlier results on minimum norm problems.

Definition. Let f be a linear functional defined on a subspace M of a vector space X. A linear functional F is said to be an *extension* of f if F is defined on a subspace N which properly contains M, and if, on M, F is identical with f. In this case, we say that F is an extension of f from M to N.

In simple terms, the Hahn-Banach theorem states that a bounded linear functional f defined only on a subspace M of a normed space can be extended to a bounded linear functional F defined on the entire space and with norm equal to the norm of f on M; i.e.,

$$\|F\| = \|f\|_M = \sup_{m \in M} \frac{|f(m)|}{\|m\|}.$$

Actually, we are able to prove a somewhat more general version of the theorem, replacing norms with sublinear functionals. This generalization is used later to prove the geometric form of the Hahn-Banach theorem.

Definition. A real-valued function p defined on a real vector space X is said to be a *sublinear functional* on X if

1. $p(x_1 + x_2) \leq p(x_1) + p(x_2)$, for all $x_1, x_2 \in X$.
2. $p(\alpha x) = \alpha p(x)$, for all $\alpha \geq 0$ and $x \in X$.

Obviously, any norm is a sublinear functional.

Theorem 1. (*Hahn-Banach Theorem, Extension Form*) *Let X be a real linear normed space and p a continuous sublinear functional on X. Let f be a linear functional defined on a subspace M of X satisfying $f(m) \leq p(m)$ for all $m \in M$. Then there is an extension F of f from M to X such that $F(x) \leq p(x)$ on X.*

Proof. The theorem is true in an arbitrary normed linear space, but our proof assumes that X is separable. The general result, however, is obtained by exactly the same method together with a simple application of Zorn's lemma. The reader familiar with Zorn's lemma should have little difficulty generalizing the proof. The basic idea is to extend f one dimension at a time and then apply induction.

Suppose y is a vector in X not in M. Consider all elements of the subspace $[M + y]$. Such an element x has a unique representation of the form $x = m + \alpha y$, where $m \in M$ and α is a real scalar. An extension g of f from M to $[M + y]$ has the form

$$g(x) = f(m) + \alpha g(y)$$

and, hence, the extension is specified by prescribing the constant $g(y)$. We must show that this constant can be chosen so that $g(x) \leq p(x)$ on $[M + y]$.

For any two elements m_1, m_2 in M, we have

$$f(m_1) + f(m_2) = f(m_1 + m_2) \leq p(m_1 + m_2) \leq p(m_1 - y) + p(m_2 + y)$$

or

$$f(m_1) - p(m_1 - y) \leq p(m_2 + y) - f(m_2),$$

and hence

$$\sup_{m \in M} [f(m) - p(m - y)] \leq \inf_{m \in M} [p(m + y) - f(m)].$$

Therefore, there is a constant c such that

$$\sup_{m \in M} [f(m) - p(m - y)] \leq c \leq \inf_{m \in M} [p(m + y) - f(m)].$$

For the vector $x = m + \alpha y \in [M + y]$, we define $g(x) = f(m) + \alpha c$. We must show that $g(m + \alpha y) \leq p(m + \alpha y)$.

If $\alpha > 0$, then

$$\alpha c + f(m) = \alpha\left[c + f\left(\frac{m}{\alpha}\right)\right] \leq \alpha\left[p\left(\frac{m}{\alpha} + y\right) - f\left(\frac{m}{\alpha}\right) + f\left(\frac{m}{\alpha}\right)\right]$$

$$= \alpha p\left(\frac{m}{\alpha} + y\right) = p(m + \alpha y)$$

If $\alpha = -\beta < 0$, then

$$-\beta c + f(m) = \beta\left[-c + f\left(\frac{m}{\beta}\right)\right] \leq \beta\left[p\left(\frac{m}{\beta} - y\right) - f\left(\frac{m}{\beta}\right) + f\left(\frac{m}{\beta}\right)\right]$$

$$= \beta p\left(\frac{m}{\beta} - y\right) = p(m - \beta y).$$

Thus $g(m + \alpha y) \leq p(m + \alpha y)$ for all α and g is an extension of f from M to $[M + y]$.

Now let $\{x_1, x_2, \ldots, x_n, \ldots\}$ be a countable dense set in X. From this set of vectors select, one at a time, a subset of vectors $\{y_1, y_2, \ldots, y_n, \ldots\}$ which is independent and independent of the subspace M. The set $\{y_1, y_2, \ldots, y_n, \ldots\}$ together with the subspace M generates a subspace S dense in X.

The functional f can be extended to a functional g on the subspace S by extending f from M to $[M + y_1]$, then to $[[M + y_1] + y_2]$, and so on.

Finally, the resulting g (which is continuous since p is) can be extended by continuity from the dense subspace S to the space X; suppose $x \in X$, then there exists a sequence $\{s_n\}$ of vectors in S converging to x. Define $F(x) = \lim_{n\to\infty} g(s_n)$. F is obviously linear and $F(x) \leftarrow g(s_n) \leq p(s_n) \rightarrow p(x)$ so $F(x) \leq p(x)$ on X. ∎

The version of the Hahn-Banach extension theorem given above is by no means the most general available. It should be noted in particular that since f and its extension F are dominated by the continuous sublinear functional p, both f and F are continuous linear functionals. A more general version of the theorem requires X only to be a linear vector space and, hence, neither continuity of the functionals nor separability of the space plays a role. Neither of these restrictions is of practical importance to us; in all applications considered in this book, the linear functions are bounded and the Hahn-Banach theorem is applied only to separable normed spaces, although the dual spaces may be nonseparable.

Corollary 1. *Let f be a bounded linear functional defined on a subspace M of a real normed vector space X. Then there is a bounded linear functional F defined on X which is an extension of f and which has norm equal to the norm of f on M.*

Proof. Take $p(x) = \|f\|_M \|x\|$ in the Hahn-Banach Theorem. ∎

The following corollary establishes the existence of nontrivial bounded linear functionals on an arbitrary normed space.

Corollary 2. *Let x be an element of a normed space X. Then there is a nonzero bounded linear functional F on X such that $F(x) = \|F\|\|x\|$.*

Proof. Assume $x \neq \theta$. On the one-dimensional subspace generated by x, define $f(\alpha x) = \alpha\|x\|$. Then f is a bounded linear functional with norm unity which, by Corollary 1, can be extended to a bounded linear functional F on X with norm unity. This functional satisfies the requirements.

If $x = \theta$, any bounded linear functional (existence of one is now established) will do. ∎

The converse of Corollary 2 is not generally true even in Banach space, as the following example illustrates.

Example 1. Let $X = l_1$, $X^* = l_\infty$. For $x = \{\xi_1, \xi_2, \xi_3, \ldots\} \in X$, we define

$$f(x) = \sum_{i=1}^{\infty} \left(1 - \frac{1}{i}\right)\xi_i.$$

It is clear that $f \in X^*$, $\|f\| = 1$. However, the reader may verify by elementary analysis that $f(x) < \|x\|$ for all nonzero $x \in l_1$.

The Hahn-Banach theorem, particularly Corollary 1, is perhaps most profitably viewed as an existence theorem for a minimization problem. Given an f on a subspace M of a normed space, it is not difficult to extend f to the whole space. An arbitrary extension, however, will in general be unbounded or have norm greater than the norm of f on M. We therefore pose the problem of selecting the extension of minimum norm. The Hahn-Banach theorem both guarantees the existence of a minimum norm extension and tells us the norm of the best extension.

5.5 The Dual of $C[a, b]$

The Hahn-Banach theorem is a useful tool for many problems in classical analysis as well as for optimization problems. As an example of its use, we characterize the dual space of $C[a, b]$. This result is of considerable interest for applications since many problems are naturally formulated on $C[a, b]$.

Theorem 1. (*Riesz Representation Theorem*) *Let f be a bounded linear functional on $X = C[a, b]$. Then there is a function v of bounded variation on $[a, b]$ such that for all $x \in X$*

$$f(x) = \int_a^b x(t)\, dv(t)$$

and such that the norm of f is the total variation of v on $[a, b]$. Conversely, every function of bounded variation on $[a, b]$ defines a bounded linear functional on X in this way.

Proof. Let B be the space of bounded functions on $[a, b]$ with the norm of an element $x \in B$ defined as $\|x\|_B = \sup_{a \leq t \leq b} |x(t)|$. The space $C[a, b]$

can be considered as a subspace of B. Thus, if f is a bounded linear functional on $X = C[a, b]$, there is, by the Hahn-Banach theorem, a linear functional F on B which is an extension of f and has the same norm.

For any $s \in [a, b]$, define the function u_s by $u_a = 0$, and by

$$u_s(t) = \begin{cases} 1 & \text{if } a \leq t \leq s \\ 0 & \text{if } s < t \leq b \end{cases}$$

for $a < s \leq b$. Obviously, each $u_s \in B$.

We define $v(s) = F(u_s)$ and show that v is of bounded variation on $[a, b]$. For this purpose, let $a = t_0 < t_1, t_2, \ldots, < t_n = b$ be a finite partition of $[a, b]$. Denoting $\varepsilon_i = \operatorname{sgn}[v(t_i) - v(t_{i-1})]$, we may write

$$\begin{aligned} \sum_{i=1}^{n} |v(t_i) - v(t_{i-1})| &= \sum_{i=1}^{n} \varepsilon_i [v(t_i) - v(t_{i-1})] \\ &= \sum_{i=1}^{n} \varepsilon_i [F(u_{t_i}) - F(u_{t_{i-1}})] \\ &= F\left[\sum_{i=1}^{n} \varepsilon_i (u_{t_i} - u_{t_{i-1}}) \right]. \end{aligned}$$

Thus,

$$\sum_{i=1}^{n} |v(t_i) - v(t_{i-1})| \leq \|F\| \left\| \sum_{i=1}^{n} \varepsilon_i (u_{t_i} - u_{t_{i-1}}) \right\| = \|f\|$$

since

$$\|F\| = \|f\| \qquad \text{and} \qquad \left\| \sum_{i=1}^{n} \varepsilon_i (u_{t_i} - u_{t_{i-1}}) \right\| = 1$$

and hence v is of bounded variation with $\text{T.V.}(v) \leq \|f\|$.

Next we derive a representation for f on X. If $x \in X$, let

$$z(t) = \sum_{i=1}^{n} x(t_{i-1})[u_{t_i}(t) - u_{t_{i-1}}(t)]$$

where $\{t_i\}$ is again a finite partition of $[a, b]$. Then

$$\|z - x\|_B = \max_i \max_{t_{i-1} \leq t \leq t_i} |x(t_{i-1}) - x(t)|$$

which (by the uniform continuity of x) goes to zero as the partition is made arbitrarily fine. Thus, since F is continuous, $F(z) \to F(x) = f(x)$. But

$$F(z) = \sum_{i=1}^{n} x(t_{i-1})[v(t_i) - v(t_{i-1})]$$

and, by the definition of the Stieltjes integral,

$$F(z) \to \int_a^b x(t)\, dv(t).$$

Therefore,

$$f(x) = \int_a^b x(t)\, dv(t).$$

It is a standard property of the Stieltjes integral that

$$\left| \int_a^b x(t)\, dv(t) \right| \leq \|x\| \cdot \text{T.V.}(v)$$

and, hence, $\|f\| \leq \text{T.V.}(v)$. On the other hand, we have $\|f\| \geq \text{T.V.}(v)$ and, consequently, $\|f\| = \text{T.V.}(v)$.

Conversely, if v is a function of bounded variation on $[a, b]$, the functional

$$f(x) = \int_a^b x(t)\, dv(t)$$

is obviously linear. Furthermore, f is bounded since $|f(x)| \leq \|x\| \text{T.V.}(v)$. ∎

It should be noted that Theorem 1 does not claim uniqueness of the function of bounded variation v representing a given linear functional f since, for example, the functional $x(1/2)$ can be represented by a v which is zero on $[0, 1/2)$, unity on $(1/2, 1]$, and has any value between zero and unity at the point $t = 1/2$. To remove the ambiguity, we introduce the following subspace of $BV[a, b]$.

Definition. The *normalized space of functions of bounded variation* denoted $NBV[a, b]$ consists of all functions of bounded variation on $[a, b]$ which vanish at the point a and which are continuous from the right on (a, b). The norm of an element v in this space is $\|v\| = \text{T.V.}(v)$.

With the above definition the association between the dual of $C[a, b]$ and $NBV[a, b]$ is unique. However, this normalization is not necessary in most applications, since when dealing with a specific functional, usually any representation is adequate.

5.6 The Second Dual Space

Let $x^* \in X^*$. We often employ the notation $\langle x, x^* \rangle$ for the value of the functional x^* at a point $x \in X$. Now, given $x \in X$ the equation $f(x^*) = \langle x, x^* \rangle$ defines a functional on the space X^*. The functional f defined on X^* in this way is linear since

$$f(\alpha x_1^* + \beta x_2^*) = \langle x, \alpha x_1^* + \beta x_2^* \rangle = \alpha \langle x, x_1^* \rangle + \beta \langle x, x_2^* \rangle = \alpha f(x_1^*) + \beta f(x_2^*).$$

Furthermore, since $|f(x^*)| = |\langle x, x^* \rangle| \leq \|x\| \cdot \|x^*\|$, it follows that $\|f\| \leq \|x\|$. By Corollary 2 of the Hahn-Banach theorem, there is a nonzero $x^* \in X^*$ such that $\langle x, x^* \rangle = \|x\| \|x^*\|$, so in fact $\|f\| = \|x\|$. We see then that, depending on whether x or x^* is considered fixed in $\langle x, x^* \rangle$, both X and X^* define bounded linear functionals on each other and this motivates the symmetric notation $\langle x, x^* \rangle$.

The space of *all* bounded linear functionals on X^* is denoted X^{**} and is called the *second dual* of X. The mapping $\varphi: X \to X^{**}$ defined by $x^{**} = \varphi(x)$ where $\langle x^*, x^{**} \rangle = \langle x, x^* \rangle$ is called the natural mapping of X into X^{**}. In other words, φ maps members of X into the functionals they generate on X^* through the symmetric notation. This mapping is linear and, as shown in the preceding paragraph, is norm preserving (i.e., $\|\varphi(x)\| = \|x\|$). Generally, however, the natural mapping of X into X^{**} is not onto. There may be elements of X^{**} that cannot be represented by elements in X. On the other hand, there are important cases in which the natural mapping is onto.

Definition. A normed linear space X is said to be *reflexive* if the natural mapping $\varphi: X \to X^{**}$ is onto. In this case we write $X = X^{**}$.

Example 1. The l_p and L_p spaces, $1 < p < \infty$, are reflexive since $l_p^* = l_q$ where $1/p + 1/q = 1$ and, thus, $l_p^{**} = l_q^* = l_p$.

Example 2. l_1 and L_1 are not reflexive.

Example 3. Any Hilbert space is reflexive.

Reflexive spaces enjoy a number of useful properties not found in arbitrary normed spaces. For instance, the converse of Corollary 2 of the Hahn-Banach theorem holds in a reflexive space; namely, given $x^* \in X^*$ there is an $x \in X$ with $\langle x, x^* \rangle = \|x\| \|x^*\|$.

5.7 Alignment and Orthogonal Complements

In general, for any $x \in X$ and any $x^* \in X^*$ we have $\langle x, x^* \rangle \leq \|x^*\| \|x\|$. In Hilbert space we have equality in this relation if and only if the functional x^* is represented by a nonnegative multiple of x, i.e., if and only if $\langle x, x^* \rangle = (x \mid \alpha x)$ for some $\alpha \geq 0$. Motivated by the Hilbert space situation, we introduce the following definition.

Definition. A vector $x^* \in X^*$ is said to be *aligned* with a vector $x \in X$ if $\langle x, x^* \rangle = \|x^*\| \|x\|$.

Alignment is a relation between vectors in two distinct vector spaces: a normed space and its normed dual.

Example 1. Let $X = L_p[a, b], 1 < p < \infty$, and $X^* = L_q[a, b], 1/p + 1/q = 1$. The condition for two functions $x \in L_p$, $y \in L_q$ to be aligned follow directly from the conditions for equality in the Hölder inequality, namely,

$$\int_a^b x(t)y(t)\,dt = \left\{\int_a^b |x(t)|^p\,dt\right\}^{1/p} \cdot \left\{\int_a^b |y(t)|^q\,dt\right\}^{1/q}$$

if and only if $x(t) = K[\operatorname{sgn} y(t)]|y(t)|^{q/p}$ for some constant K.

Example 2. Let $x \in X = C[a, b]$ and let Γ be the set of points $t \in [a, b]$ at which $|x(t)| = \|x\|$. In general, Γ may be infinite or finite but it is always nonempty. A bounded linear functional $x^*(x) = \int_a^b x(t)\,dv(t)$ is aligned with x if and only if v varies only on Γ and v is nondecreasing at t if $x(t) > 0$ and nonincreasing if $x(t) < 0$. (We leave the details to the reader.) Thus, if Γ is finite, an aligned functional must consist of a finite number of step discontinuities. See Figure 5.1

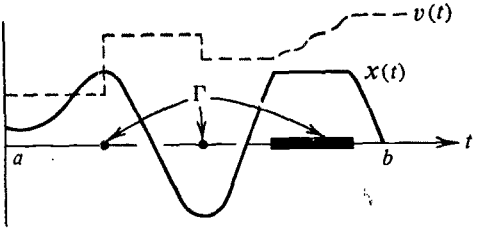

Figure 5.1 Aligned functions

The notion of orthogonality can be introduced in normed spaces through the dual space.

Definition. The vectors $x \in X$ and $x^* \in X^*$ are said to be *orthogonal* if $\langle x, x^* \rangle = 0$.

Since the dual of a Hilbert space X is itself X, in the sense described in Section 5.3 by the Riesz-Fréchet theorem, the definition of orthogonality given above can be regarded as a generalization of the corresponding Hilbert space definition.

Definition. Let S be a subset of a normed linear space X. The *orthogonal complement*[1] of S, denoted $S^\perp$, consists of all elements $x^* \in X^*$ orthogonal to every vector in S.

Given a subset U of the dual space X^*, its orthogonal complement $U^\perp$ is in X^{**}. A more useful concept in this case, however, is described below.

[1] The term *annihilator* is often used in place of orthogonal complement.

Definition. Given a subset U of the dual space X^*, we define the *orthogonal complement of U in X* as the set ${}^{\perp}U \subset X$ consisting of all elements in X orthogonal to every vector in U.

The set ${}^{\perp}U$ may be thought of as the intersection of $U^{\perp}$ with X, where X is considered to be imbedded in X^{**} by the natural mapping. Many of the relations among orthogonal complements for Hilbert space generalize to normed spaces. In particular we have the following fundamental duality result.

Theorem 1. *Let M be a closed subspace of a normed space X. Then* ${}^{\perp}[M^{\perp}] = M$.

Proof. It is clear that $M \subset {}^{\perp}[M^{\perp}]$. To prove the converse, let $x \notin M$. On the subspace $[x + M]$ generated by x and M, define the linear functional $f(\alpha x + m) = \alpha$ for $m \in M$. Then

$$\|f\| = \sup_{m \in M} \frac{f(x+m)}{\|x+m\|} = \frac{1}{\inf_m \|x+m\|}$$

and since M is closed, $\|f\| < \infty$. Thus by the Hahn-Banach theorem, we can extend f to an $x^* \in X^*$. Since f vanishes on M, we have $x^* \in M^{\perp}$. But also $\langle x, x^* \rangle = 1$ and thus $x \notin {}^{\perp}[M^{\perp}]$. ∎

5.8 Minimum Norm Problems

In this section we consider the question of determining a vector in a subspace M of a normed space which best approximates a given vector x in the sense of minimum norm. This section thus extends the results of Chapter 3 for minimum norm problems in Hilbert space.

We recall that if M is a closed subspace in Hilbert space, there is always a unique solution to the minimum norm problem and the solution satisfies an orthogonality condition. Furthermore, the projection theorem leads to a linear equation for determining the unknown optimizing vector. Even limited experience with nonquadratic optimization problems warns us that the situation is likely to be more complex in arbitrary normed spaces. The optimal vector, if it exists, may not be unique and the equations for the optimal vector will generally be nonlinear. Nevertheless, despite these difficulties, we find that the theorems of Chapter 3 have remarkable analogs here. As before, the key concept is that of orthogonality and our principal result is an analog of the projection theorem.

As an example of the difficulties encountered in arbitrary normed space, we consider a simple two-dimensional minimum norm problem that does not have a unique solution.

Example 1. Let X be the space of pairs of real numbers $x = (\xi_1, \xi_2)$ with $\|x\| = \max_{i=1,2} |\xi_i|$. Let M be the subspace of X consisting of all those vectors having their second component zero, and consider the fixed point $x = (2, 1)$. The minimum distance from x to M is obviously 1, but any vector in M of the form $m = (a, 0)$ where $1 \leq a \leq 3$ satisfies $\|x - m\| = 1$. The situation is sketched in Figure 5.2.

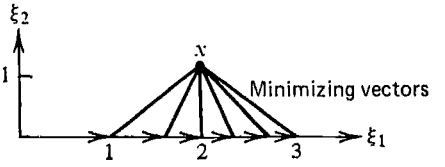

Figure 5.2 Solution to Example 1

The following two theorems essentially constitute a complete resolution of minimum norm problems of this kind. These theorems, when specialized to Hilbert space, contain all the conclusions of the projection theorem except the uniqueness of the solution. When uniqueness holds, however, it is fairly easy to prove separately.

Furthermore, the following two theorems contain even more information than the projection theorem. They introduce a duality principle stating the equivalence of two extremization problems: one formulated in a normed space and the other in its dual. Often the transition from one problem to its dual results in significant simplification or enhances physical and mathematical insight. Some infinite-dimensional problems can be converted to equivalent finite-dimensional problems by consideration of the dual problem.

Theorem 1. *Let x be an element in a real normed linear space X and let d denote its distance from the subspace M. Then,*

$$d = \inf_{m \in M} \|x - m\| = \max_{\substack{\|x^*\| \leq 1 \\ x^* \in M^\perp}} \langle x, x^* \rangle \tag{1}$$

where the maximum on the right is achieved for some $x_0^* \in M^\perp$.

If the infimum on the left is achieved for some $m_0 \in M$, then x_0^* is aligned with $x - m_0$.

Proof. For $\varepsilon > 0$, let $m_\varepsilon \in M$ satisfy $\|x - m_\varepsilon\| \leq d + \varepsilon$. Then for any $x^* \in M^\perp$, $\|x^*\| \leq 1$, we have

$$\langle x, x^* \rangle = \langle x - m_\varepsilon, x^* \rangle \leq \|x^*\| \|x - m_\varepsilon\| \leq d + \varepsilon.$$

Since ε was arbitrary, we conclude that $\langle x, x^* \rangle \leq d$. Therefore, the proof of the first part of the theorem is complete if we exhibit any x_0^* for which $\langle x, x_0^* \rangle = d$.

Let N be the subspace $[x + M]$. Elements of N are uniquely representable in the form $n = \alpha x + m$, with $m \in M$, α real. Define the linear functional f on N by the equation $f(n) = \alpha d$. We have

$$\|f\| = \sup_N \frac{|f(n)|}{\|n\|} = \sup \frac{|\alpha| d}{\|\alpha x + m\|}$$

$$= \sup \frac{|\alpha| d}{|\alpha| \left\| x + \frac{m}{\alpha} \right\|} = \frac{d}{\inf \left\| x + \frac{m}{\alpha} \right\|} = 1.$$

Now form the Hahn-Banach extension, x_0^*, of f from N to X. Then $\|x_0^*\| = 1$ and $x_0^* = f$ on N. By construction, we have $x_0^* \in M^\perp$ and $\langle x, x_0^* \rangle = d$; hence x_0^* satisfies the requirements of the first part of the theorem.

Now assume that there is an $m_0 \in M$ with $\|x - m_0\| = d$ and let x_0^* be any element such that $x_0^* \in M^\perp$, $\|x_0^*\| = 1$, and $\langle x, x_0^* \rangle = d$—the x_0^* constructed above being one possibility. Then

$$\langle x - m_0, x_0^* \rangle = \langle x, x_0^* \rangle = d = \|x_0^*\| \|x - m_0\|$$

and x_0^* is aligned with $x - m_0$. ∎

The reader should attempt to visualize the problem and the relation (1) geometrically. The theorem becomes quite clear if we imagine that the error $x - m_0$ is orthogonal to M.

Theorem 1 states the equivalence of two optimization problems: one in X called the primal problem and the other in X^* called the dual problem. The problems are related through both the optimal values of their respective objective functionals and an alignment condition on their solution vectors. Since in many spaces alignment can be explicitly characterized, the solution of either problem often leads directly to the solution of the other. Duality relations such as this are therefore often of extreme practical as well as theoretical significance in optimization problems.

If we take only a portion of the above theorem, we obtain a generalization of the projection theorem.

Corollary 1. *Let x be an element of a real normed linear vector space X and let M be a subspace of X. A vector $m_0 \in M$ satisfies $\|x - m_0\| \leq \|x - m\|$ for all $m \in M$ if and only if there is a nonzero vector $x^* \in M^\perp$ aligned with $x - m_0$.*

Proof. The "only if" part follows directly from Theorem 1. To prove the "if" part, assume that $x - m_0$ is aligned with $x^* \in M^\perp$. Without loss of generality, take $\|x^*\| = 1$. For all $m \in M$ we have

$$\langle x, x^* \rangle = \langle x - m, x^* \rangle \leq \|x - m\|$$

whereas

$$\langle x, x^* \rangle = \langle x - m_0, x^* \rangle = \|x - m_0\|.$$

Thus, $\|x - m_0\| \leq \|x - m\|$. ∎

As a companion to Theorem 1, we have:

Theorem 2. *Let M be a subspace in a real normed space X. Let $x^* \in X^*$ be a distance d from $M^\perp$. Then*

$$d = \min_{m^* \in M^\perp} \|x^* - m^*\| = \sup_{\substack{x \in M \\ \|x\| \leq 1}} \langle x, x^* \rangle \tag{2}$$

where the minimum on the left is achieved for $m_0^* \in M^\perp$. If the supremum on the right is achieved for some $x_0 \in M$, then $x^* - m_0^*$ is aligned with x_0.

Proof. We denote the right-hand side of relation (2) by $\|x^*\|_M$ because it is the norm of the functional x^* restricted to the subspace M. For any $m^* \in M^\perp$, we have

$$\begin{aligned} \|x^* - m^*\| &= \sup_{\|x\| \leq 1} [\langle x, x^* \rangle - \langle x, m^* \rangle] \geq \sup_{\substack{x \in M \\ \|x\| \leq 1}} [\langle x, x^* \rangle - \langle x, m^* \rangle] \\ &= \sup_{\substack{x \in M \\ \|x\| \leq 1}} \langle x, x^* \rangle = \|x^*\|_M. \end{aligned}$$

Thus, $\|x^* - m^*\| \geq \|x^*\|_M$ and the first part of the theorem is proved if an $m_0^* \in M^\perp$ can be found giving equality.

Consider x^* restricted to the subspace M. The norm of x^* so restricted is $\|x^*\|_M$. Let y^* be the Hahn-Banach extension to the whole space of the restriction of x^*. Thus, $\|y^*\| = \|x^*\|_M$ and $x^* - y^* = 0$ on M. Put $m_0^* = x^* - y^*$. Then $m_0^* \in M^\perp$ and $\|x^* - m_0^*\| = \|x^*\|_M$.

If the supremum on the right is achieved by some $x_0 \in M$, then obviously $\|x_0\| = 1$ and $\|x^* - m_0^*\| = \langle x_0, x^* \rangle = \langle x_0, x^* - m_0^* \rangle$. Thus $x^* - m_0^*$ is aligned with x_0. ∎

Theorem 2 guarantees the existence of a solution to the minimum norm problem if the problem is appropriately formulated in the dual of a normed space. This result simply reflects the fact that the Hahn-Banach theorem establishes the existence of certain linear functionals rather than vectors and establishes the general rule, which we adhere to in applications, that minimum norm problems must be formulated in a dual space if one is to guarantee the existence of a solution.

The duality of the above theorems can be displayed more explicitly in terms of some fairly natural notation. Assuming M to be a closed subspace, we write

$$\|x\|_M = \inf_{m \in M} \|x - m\|$$

since this is the norm of the coset generated by x in the space X/M. For an element $x^* \in X^*$, we write, as in the proof of Theorem 2,

$$\|x^*\|_M = \sup_{\substack{\|x\| \leq 1 \\ x \in M}} \langle x, x^* \rangle$$

since this is the norm of x^* considered as a functional on M. In this notation, equations (1) and (2) become

$$\|x\|_M = \|x\|_{M^\perp} \tag{3}$$

$$\|x^*\|_{M^\perp} = \|x^*\|_M \tag{4}$$

where on the right side of (3) the vector x is regarded as a functional on X^*.

5.9 Applications

In this section we present examples of problem solution by use of the theory developed in the last section. Three basic guidelines are applied to our analyses of the problems considered: (1) In characterizing optimum solutions, use the alignment properties of the space and its dual. In the L_p and l_p spaces, for instance, this amounts to the conditions for equality in the Hölder inequality. (2) Try to guarantee the existence of a solution by formulating minimum norm problems in a dual space. (3) Look at the dual problem to see if it is easier than the original problem. The dual may have lower dimension or be more transparent.

Example 1. (Chebyshev Approximation) Let f be a continuous function on an interval $[a, b]$ of the real line. Suppose we seek the polynomial p of degree n (or less) that best approximates f in the sense of minimizing $\max_{a \leq t \leq b} |f(t) - p(t)|$, i.e., minimizing the maximum deviation of the two functions. In the Banach space $X = C[a, b]$, this problem is equivalent to finding the p in the $n + 1$-dimensional subspace N of n-th degree polynomials that minimizes $\|f - p\|$. We may readily establish the existence of an optimal p, say p_0, since the subspace N is finite dimensional.

Suppose $\|f - p_0\| = d > 0$, and let Γ be the set of points t in $[a, b]$ for which $|f(t) - p_0(t)| = d$. We show that Γ contains at least $n + 2$ points.

According to Theorem 1 of Section 5.8, the optimal solution p_0 must be such that $f - p_0$ is aligned with an element in $N^\perp \subset X^* = NBV[a, b]$. Assume that Γ contained $m < n + 2$ points t_i: $a \leq t_1 < t_2 < \cdots < t_m \leq b$. If $v \in NBV[a, b]$ is aligned with $f - p_0$, v varies only on these points (see Example 2, Section 5.7), and hence must consist of jump discontinuities at the t_i's. Let t_k be a point of jump discontinuity of v. Then the polynomial $q(t) = \prod_{i \neq k} (t - t_i)$ is in N but has $\int_a^b q \, dv \neq 0$ and consequently, $v \notin N^\perp$.

Therefore, $f - p_0$ is not aligned with any nonzero element of $N^\perp$ and hence Γ must contain at least $n + 2$ points. We have therefore proved the classic result of Tonelli: *If f is continuous on $[a, b]$ and p_0 is the polynomial of degree n (or less) minimizing* $\max_{t \in [a, b]} |f(t) - p(t)|$, *then* $|f(t) - p_0(t)|$ *achieves its maximum at at least $n + 2$ points in $[a, b]$.*

Many problems amenable to the theory of the last section are most naturally formulated as finding the vector of minimum norm in a linear variety rather than as finding the best approximation on a subspace. A standard problem of this kind arising in several contexts is to find an element of minimum norm satisfying a finite number of linear constraints. To guarantee existence of a solution, we consider the unknown x^* in a dual space X^* and express the constraints in the form

$$\begin{aligned} \langle y_1, x^* \rangle &= c_1 \\ \langle y_2, x^* \rangle &= c_2 \\ &\vdots \quad\quad \vdots \\ \langle y_n, x^* \rangle &= c_n, \qquad y_i \in X. \end{aligned}$$

If $\bar{x}^*$ is any vector satisfying the constraints, we have

$$d = \min_{\langle y_i, x^* \rangle = c_i} \|x^*\| = \min_{m^* \in M^\perp} \|\bar{x}^* - m^*\|$$

where M denotes the space generated by the y_i's. From Theorem 2, Section 5.8, this becomes

$$d = \min_{m^* \in M^\perp} \|\bar{x}^* - m^*\| = \sup_{\substack{x \in M \\ \|x\| \le 1}} \langle x, \bar{x}^* \rangle.$$

Any vector in M is of the form $x = \sum_{i=1}^n a_i y_i$ or, symbolically, Ya; thus, since M is finite dimensional,

$$d = \min_{\langle y_i, x^* \rangle = c_i} \|x^*\| = \max_{\|Ya\| \le 1} \langle Ya, \bar{x}^* \rangle = \max_{\|Ya\| \le 1} c'a,$$

the last equality following from the fact that $\bar{x}^*$ satisfies the constraints. The quantity $c'a$ denotes the usual inner product of the two n-vectors a and c with components a_i and c_i. Furthermore, we have the alignment properties of Theorem 2, Section 5.8, and thus the following corollary.

Corollary 1. *Let $y_i \in X$, $i = 1, 2, \ldots, n$, and suppose the system of linear equalities $\langle y_i, x^* \rangle = c_i$, $i = 1, 2, \ldots, n$ is consistent; i.e., the set*

$$D = \{x^* \in X^* : \langle y_i, x^* \rangle = c_i,\ i = 1, 2, \ldots, n\}$$

is nonempty. Then

$$\min_{x^* \in D} \|x^*\| = \max_{\|Ya\| \leq 1} c'a.$$

Furthermore, the optimal x^* is aligned with the optimal Ya.

Example 2. (A Control Problem) Consider the problem of selecting the field current $u(t)$ on $[0, 1]$ to drive a motor governed by

$$\ddot{\theta}(t) + \dot{\theta}(t) = u(t)$$

from the initial conditions $\theta(0) = \dot{\theta}(0) = 0$ to $\theta(1) = 1$, $\dot{\theta}(1) = 0$ in such a way as to minimize $\max_{0 \leq t \leq 1} |u(t)|$. This example is similar to Example 1, Section 3.10, but now our objective function reflects a concern with possible damage due to excessive current rather than with total energy.

The problem can be thought of as being formulated in $C[0, 1]$, but since this is not the dual of any normed space, we are not guaranteed that a solution exists in $C[0, 1]$. Thus, instead we take $X = L_1[0, 1]$, $X^* = L_\infty[0, 1]$ and seek $u \in X^*$ of minimum norm. The constraints are, as in Example 1, Section 3.10,

$$\int_0^1 e^{(t-1)} u(t)\, dt = 0$$

$$\int_0^1 \{1 - e^{(t-1)}\} u(t)\, dt = 1.$$

From Corollary 1,

$$\min \|u\| = \max_{\|a_1 y_1 + a_2 y_2\| \leq 1} a_2.$$

The norm on the right is the norm in $X = L_1[0, 1]$. Thus, the two constants a_1, a_2 must satisfy

$$\int_0^1 |(a_1 - a_2)e^{(t-1)} + a_2|\, dt \leq 1.$$

Maximization of a_2 subject to this constraint is a straightforward task, but we do not carry out the necessary computations. Instead we show that the general nature of the optimal control is easily deduced from the alignment requirement. Obviously, the function $a_1 y_1(t) + a_2 y_2(t)$, being the sum of a constant and an exponential term, can change sign at most once, and since the optimal u is aligned with this function, u must be "bang-bang" (i.e., it must have values $\pm M$ for some M) and changes sign at most once. We leave it to the reader to verify this by characterizing alignment between $L_1[0, 1]$ and $L_\infty[0, 1]$.

Example 3. (Rocket Problem) Consider the problem of selecting the thrust program $u(t)$ for a vertically ascending rocket-propelled vehicle, subject only to the forces of gravity and rocket thrust in order to reach a given altitude with minimum fuel expenditure. Assuming fixed unit mass, unit gravity, and zero initial conditions, the altitude $x(t)$ is governed by a differential equation of the form

$$\ddot{x}(t) = u(t) - 1 \qquad x(0) = \dot{x}(0) = 0.$$

This equation can be integrated twice (once by parts) to give

$$x(T) = \int_0^T (T - t)u(t)\, dt - \frac{T^2}{2}.$$

Our problem is to attain a given altitude, say $x(T) = 1$, while minimizing the fuel expense.

$$\int_0^T |u(t)|\, dt.$$

The final time T is in general unspecified, but we approach the problem by finding the minimum fuel expenditure for each fixed T and then minimizing over T.

For a fixed T the optimization problem reduces to that of finding u minimizing

$$\int_0^T |u(t)|\, dt$$

while satisfying the single linear constraint

$$\int_0^T (T - t)u(t)\, dt = 1 + \frac{T^2}{2}.$$

At first sight we might regard this problem as one in $L_1[0, T]$. Since, however, $L_1[0, T]$ is not the dual of any normed space, we imbed our problem in the space $NBV[0, T]$ and associate control elements u with the derivatives of elements v in $NBV[0, T]$. Thus the problem becomes that of finding the $v \in NBV[0, 1]$ minimizing

$$\int_0^T |dv(t)| = \text{T.V.}(v) = \|v\|$$

subject to

$$\int_0^T (T - t)\, dv(t) = 1 + \frac{T^2}{2}.$$

According to Corollary 1,

$$\min \|v\| = \max_{\|(T-t)a\| \le 1} \left[a\left(1 + \frac{T^2}{2}\right) \right],$$

which is only a one-dimensional problem. The norm on the right-hand side is taken in $C[0, T]$, the space to which $NBV[0, T]$ is dual. In $C[0, T]$ we have

$$\|(T-t)a\| = \max_{0 \le t \le T} |(T-t)a| = T\,|a|$$

since the maximum occurs at $t = 0$. Thus the optimal choice is $a = 1/T$ and

$$\min \|v\| = \left(1 + \frac{T^2}{2}\right) \frac{1}{T}.$$

The optimal v must be aligned with $(T-t)a$ and, hence, can vary only at $t = 0$. Therefore, we conclude that v is a step function and u is an impulse (or delta function) at $t = 0$. The best final time can be obtained by differentiating the optimal fuel expenditure with respect to T. This leads to the final result

$$T = \sqrt{2}$$
$$\min \|v\| = \sqrt{2}$$

and

$$v = \begin{cases} 0 & t = 0 \\ \sqrt{2} & 0 < t \le \sqrt{2}. \end{cases}$$

Note that our early observation that the problem should be formulated in $NBV[0, T]$ rather than $L_1[0, T]$ turned out to be crucial since the optimal u is an impulse.

*5.10 Weak Convergence

An interesting and important concept that arises naturally upon the introduction of the dual space is that of weak convergence. It is important for certain problems in analysis and plays an indirect role in many optimization problems.

Definition. A sequence $\{x_n\}$ in a normed linear vector space X is said to *converge weakly* to $x \in X$ if for every $x^* \in X^*$ we have $\langle x_n, x^* \rangle \to \langle x, x^* \rangle$. In this case we write $x_n \to x$ weakly.

Our earlier notion of convergence, convergence in norm, is sometimes referred to as *strong convergence*. We have the following result.

Proposition 1. *If $x_n \to x$ strongly, then $x_n \to x$ weakly.*

Proof. $|\langle x_n, x^* \rangle - \langle x, x^* \rangle| \leq \|x^*\| \|x_n - x\| \to 0$. ∎

There are, however, sequences that converge weakly but not strongly.

Example 1. In $X = l_2$ consider the elements $x_n = \{0, 0, \ldots, 0, 1, 0, \ldots\}$ with the 1 in the n-th place. For any $y = \{\eta_1, \eta_2, \ldots\} \in l_2 = X^*$, we have $(x_n | y) = \eta_n \to 0$ as $n \to \infty$. Thus $x_n \to \theta$ weakly. However, $x_n \not\to \theta$ strongly since $\|x_n\| = 1$.

Starting with a normed space X, we form X^* and define weak convergence on X in terms of X^*. The same technique can be applied to X^* with weak convergence being defined in terms of X^{**}. However, there is a more important notion of convergence in X^* defined in terms of X rather than X^{**}.

Definition. A sequence $\{x_n^*\}$ in X^* is said to *converge weak-star* (or *weak**) to the element x^* if for every $x \in X$, $\langle x, x_n^* \rangle \to \langle x, x^* \rangle$. In this case we write $x_n^* \to x^*$ weak*.

Thus in X^* we have three separate notions of convergence: strong, weak, and weak*; furthermore, strong implies weak, and weak implies weak* convergence. In general, weak* convergence does not imply weak convergence in X^*.

Example 2. Let $X = c_0$, the space of infinite sequences convergent to zero, with norm equal to the maximum absolute value of the terms. Then $X^* = l_1$, $X^{**} = l_\infty$. (See Section 5.3.) In $X^* = l_1$, let $x_n^* = \{0, 0, 0, \ldots, 0, 1, 0, 0, 0, \ldots\}$, the term 1 being at the n-th place. Then $x_n^* \to \theta$ weak* but $x_n^* \not\to \theta$ weakly since $\langle x_n^*, x^{**} \rangle \not\to 0$ for $x^{**} = \{1, 1, 1, \ldots\}$.

It was shown in Section 2.13 that a continuous functional on a compact set achieves a maximum and a minimum. However, the usual definition of compactness, i.e., compactness with respect to strong convergence, is so severe that this property can be used only in very special circumstances such as in finite-dimensional spaces. Weak compactness and weak* compactness are less severe requirements on a set; indeed, such compactness criteria (especially weak*) provide alternative explanations for the existence of solutions to optimization problems.

Definition. A set $K \subset X^*$ is said to be *weak* compact* if every infinite sequence from K contains a weak* convergent subsequence.

Theorem 1. *(Alaoglu) Let X be a real normed linear space. The closed unit sphere in X^* is weak* compact.*

Proof. Although the theorem is true in general, we only prove it for the case where X is separable (although X^* need not be separable), which is adequate for all of the examples treated in this book.

Let $\{x_n^*\}$ be an infinite sequence in X^* such that $\|x_n^*\| \leq 1$. Let $\{x_k\}$ be a sequence from X dense in X. The sequence $\{\langle x_1, x_n^*\rangle\}$ of real numbers is bounded and thus contains a convergent subsequence which we denote $\{\langle x_1, x_{n1}^*\rangle\}$. Likewise the sequence $\{\langle x_2, x_{n1}^*\rangle\}$ contains a convergent subsequence $\{\langle x_2, x_{n2}^*\rangle\}$. Continuing in this fashion to extract subsequences $\{\langle x_k, x_{nk}^*\rangle\}$, we then form the diagonal sequence $\{x_{nn}^*\}$ in X^*.

The sequence $\{x_{nn}^*\}$ converges on the dense subset $\{x_k\}$ of X; i.e., the sequence of real numbers $\{\langle x_k, x_{nn}^*\rangle\}$ converges to a real number for each x_k. The proof of the theorem is complete if we show that $\{x_{nn}^*\}$ converges weak* to an element $x^* \in X^*$.

Fix $x \in X$ and $\varepsilon > 0$. Then for any n, m, k,

$$\begin{aligned}|\langle x, x_{nn}^*\rangle - \langle x, x_{mm}^*\rangle| &\leq |\langle x, x_{nn}^*\rangle - \langle x_k, x_{nn}^*\rangle| + |\langle x_k, x_{nn}^*\rangle - \langle x_k, x_{mm}^*\rangle| \\ &\quad + |\langle x_k, x_{mm}^*\rangle - \langle x, x_{mm}^*\rangle| \\ &\leq 2\|x_k - x\| + |\langle x_k, x_{nn}^*\rangle - \langle x_k, x_{mm}^*\rangle|.\end{aligned}$$

Now choose k so that $\|x_k - x\| < \varepsilon/3$ and then N so that for $n, m > N$, $|\langle x_k, x_{nn}^*\rangle - \langle x_k, x_{mm}^*\rangle| < \varepsilon/3$. Then for $n, m > N$, we have $|\langle x, x_{nn}^*\rangle - \langle x, x_{mm}^*\rangle| < \varepsilon$. Thus, $\{\langle x, x_{nn}^*\rangle\}$ is a Cauchy sequence and consequently converges to a real number $\langle x, x^*\rangle$.

The functional $\langle x, x^*\rangle$ so defined is clearly linear and has $\|x^*\| \leq 1$. ∎

Definition. A functional (possibly nonlinear) defined on a normed space X is said to be *weakly continuous* at x_0 if given $\varepsilon > 0$ there is a $\delta > 0$ and a finite collection $\{x_1^*, x_2^*, \ldots, x_n^*\}$ from X^* such that $|f(x) - f(x_0)| < \varepsilon$ for all x such that $|\langle x - x_0, x_i^*\rangle| < \delta$ for $i = 1, 2, \ldots, n$. *Weak* continuity* of a functional defined on X^* is defined analogously with the roles of X and X^* interchanged.

The reader can easily verify that the above definition assures that if f is weakly continuous, $x_n \to x$ weakly implies that $f(x_n) \to f(x)$. We also leave it to the reader to prove the following result.

Theorem 2. *Let f be a weak* continuous real-valued functional on a weak* compact subset S of X^*. Then f is bounded on S and achieves its maximum on S.*

A special application of Theorems 1 and 2 is to the problem of maximizing $\langle x, x^*\rangle$ for a fixed $x \in X$ while x^* ranges over the unit sphere in X^*.

Since the unit sphere in X^* is weak* compact and $\langle x, x^* \rangle$ is a weak* continuous functional on X^*, the maximum is achieved. This result is equivalent to Corollary 2 of the Hahn-Banach theorem, Section 5.4.

GEOMETRIC FORM OF THE HAHN-BANACH THEOREM

5.11 Hyperplanes and Linear Functionals

In the remaining sections of this chapter we generalize the results for minimum norm problems from linear varieties to convex sets. The foundation of this development is again the Hahn-Banach theorem but in the geometric rather than extension form.

There is a major conceptual difference between the approach taken in the remainder of this chapter and that taken in the preceding sections. Linear functionals, rather than being visualized as elements of a dual space, are visualized as hyperplanes generated in the primal space. This difference in viewpoint combines the relevant aspects of both the primal and the dual into a single geometric image and thereby frees our intuition of the burden of visualizing two distinct spaces.

Definition. A *hyperplane* H in a linear vector space X is a maximal proper linear variety, that is, a linear variety H such that $H \neq X$, and if V is any linear variety containing H, then either $V = X$ or $V = H$.

This definition of hyperplane is made without explicit reference to linear functionals and thus stresses the geometric interpretation of a hyperplane. Hyperplanes are intimately related to linear functionals, however, as the following three propositions demonstrate.

Proposition 1. *Let H be a hyperplane in a linear vector space X. Then there is a linear functional f on X and a constant c such that $H = \{x : f(x) = c\}$. Conversely, if f is a nonzero linear functional on X, the set $\{x : f(x) = c\}$ is a hyperplane in X.*

Proof. Let H be a hyperplane in X. Then H is the translation of a subspace M in X, say $H = x_0 + M$. If $x_0 \notin M$, then $[M + x_0] = X$, and for $x = \alpha x_0 + m$, with $m \in M$ we define $f(x) = \alpha$. Then $H = \{x : f(x) = 1\}$. If $x_0 \in M$, we take $x_1 \notin M$, $X = [M + x_1]$, $H = M$, and define for $x = \alpha x_1 + m$, $f(x) = \alpha$. Then $H = \{x : f(x) = 0\}$.

Conversely, let f be a nonzero linear functional on X and let $M = \{x : f(x) = 0\}$. It is clear that M is a subspace. Let $x_0 \in X$ with $f(x_0) = 1$. Then for any $x \in X$, $f[x - f(x)x_0] = 0$ and, hence, $x - f(x)x_0 \in M$. Thus, $X = [x_0 + M]$, and M is a maximal proper subspace. For any real c, let x_1 be any element for which $f(x_1) = c$. Then $\{x : f(x) = c\} = \{x : f(x - x_1) = 0\} = M + x_1$ which is a hyperplane. ∎

Hyperplanes that contain the origin represent a somewhat special case, but barring these it is possible to establish a unique correspondence between hyperplanes and linear functionals.

Proposition 2. *Let H be a hyperplane in a linear vector space X. If H does not contain the origin, there is a unique linear functional f on X such that $H = \{x : f(x) = 1\}$.*

Proof. By appropriate scaling, Proposition 1 guarantees the existence of at least one such functional f. Let g be any other such functional, so that in particular $H = \{x : f(x) = 1\} = \{x : g(x) = 1\}$. It is then clear that $H \subset \{x : f(x) - g(x) = 0\}$. Since the smallest subspace of X containing H is X, it follows that $f = g$. ∎

The above considerations are quite general in that they apply to hyperplanes in an arbitrary linear vector space. A hyperplane H in a normed space X must be either closed or dense in X because, since H is a maximal linear variety, either $\bar{H} = H$ or $\bar{H} = X$. For our purposes we are primarily interested in closed hyperplanes in a normed space X. These hyperplanes correspond to the bounded linear functionals on X.

Proposition 3. *Let f be a nonzero linear functional on a normed space X. Then the hyperplane $H = \{x : f(x) = c\}$ is closed for every c if and only if f is continuous.*

Proof. Suppose first that f is continuous. Let $\{x_n\}$ be a sequence from H convergent to $x \in X$. Then $c = f(x_n) \to f(x)$ and thus $x \in H$ and H is closed. Conversely, assume that $M = \{x : f(x) = 0\}$ is closed. Let $X = [x_0 + M]$ and suppose $x_n \to x$ in X. Then $x_n = \alpha_n x_0 + m_n$, $x = \alpha x_0 + m$, and letting d denote the distance of x_0 from M (which is positive since M is closed), we have $|\alpha_n - \alpha| d \leq \|x_n - x\| \to 0$ and hence $\alpha_n \to \alpha$. Also $f(x_n) = \alpha_n f(x_0) + f(m_n) = \alpha_n f(x_0) \to \alpha f(x_0) = f(x)$. Thus f is continuous on X. ∎

If f is a nonzero linear functional on a linear vector space X, we associate with the hyperplane $H = \{x : f(x) = c\}$ the four sets

$$\{x : f(x) \leq c\}, \qquad \{x : f(x) < c\}, \qquad \{x : f(x) \geq c\}, \qquad \{x : f(x) > c\}$$

called *half-spaces* determined by H. The first two of these are referred to as *negative half-spaces* determined by f and the second two as *positive half-spaces*. If f is continuous, then the half-spaces $\{x : f(x) < c\}$, $\{x : f(x) > c\}$ are open and $\{x : f(x) \leq c\}$, $\{x : f(x) \geq c\}$ are closed.

The results of this section establish a correspondence between hyperplanes and linear functionals, particularly between the closed hyperplanes and members of the dual X^*. The fact that the correspondence is unique

for hyperplanes not containing the origin suggests that virtually all concepts in which X^* plays a fundamental role can be visualized in terms of closed hyperplanes or their corresponding half-spaces.

5.12 Hyperplanes and Convex Sets

In this section we prove the geometric form of the Hahn-Banach theorem which in simplest form says that given a convex set K containing an interior point, and given a point x_0 not in $\mathring{K}$, there is a closed hyperplane containing x_0 but disjoint from $\mathring{K}$.

If K were the unit sphere, this result would follow immediately from our earlier version of the Hahn-Banach theorem since it establishes the existence of an x_0^* aligned with x_0. For every x in the interior of the unit sphere, we then have $\langle x, x_0^* \rangle \leq \|x_0^*\| \|x\| < \|x_0^*\| \|x_0\| = \langle x_0, x_0^* \rangle$ or $\langle x, x_0^* \rangle < \langle x_0, x_0^* \rangle$, which implies that the hyperplane $\{x : \langle x, x_0^* \rangle = \langle x_0, x_0^* \rangle\}$ is disjoint from the interior of the unit sphere. If we begin with an arbitrary convex set K, on the other hand, we might try to redefine the norm on X so that K, when translated so as to contain θ, would be the unit sphere with respect to this norm. The Hahn-Banach theorem could then be applied on this new normed space. This approach is in fact successful for some special convex sets. To handle the general case, however, we must use the general Hahn-Banach theorem stated in terms of sublinear functionals instead of norms.

Definition. Let K be a convex set in a normed linear vector space X and suppose θ is an interior point of K. Then the *Minkowski functional* p of K is defined on X by

$$p(x) = \inf \left\{ r : \frac{x}{r} \in K, r > 0 \right\}.$$

We note that for K equal to the unit sphere in X the Minkowski functional is $\|x\|$. In the general case, $p(x)$ defines a kind of distance from the origin to x measured with respect to K; it is the factor by which K must be expanded so as to include x. See Figure 5.3.

Lemma 1. *Let K be a convex set containing θ as an interior point. Then the Minkowski functional p of K satisfies:*

1. $\infty > p(x) \geq 0$ *for all* $x \in X$,
2. $p(\alpha x) = \alpha p(x)$ *for* $\alpha > 0$,
3. $p(x_1 + x_2) \leq p(x_1) + p(x_2)$,
4. p *is continuous*,
5. $\overline{K} = \{x : p(x) \leq 1\}$, $\mathring{K} = \{x : p(x) < 1\}$.

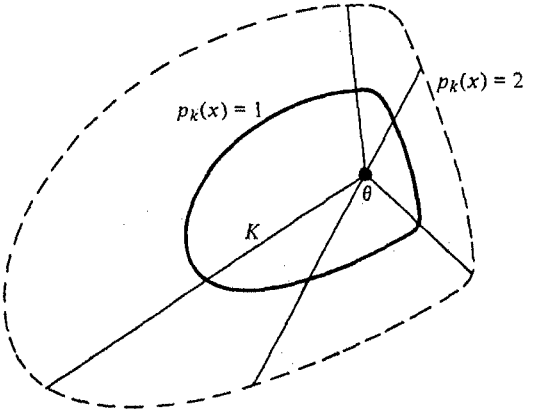

Figure 5.3 The Minkowski functional of a convex set

Proof.

1. Since K contains a sphere about θ, given x there is an $r > 0$ such that $x/r \in K$. Thus $p(x)$ is finite for all x. Obviously, $p(x) \geq 0$.
2. For $\alpha > 0$,

$$p(\alpha x) = \inf \{r: \frac{\alpha x}{r} \in K, r > 0\}$$

$$= \inf \{\alpha r': \frac{x}{r'} \in K, r' > 0\}$$

$$= \alpha \inf \{r': \frac{x}{r'} \in K, r' > 0\} = \alpha p(x).$$

3. Given x_1, x_2 and $\varepsilon > 0$, choose r_1, r_2 such that $p(x_i) < r_i < p(x_i) + \varepsilon$, $i = 1, 2$. By No. 2, $p(x_i/r_i) < 1$ and so $x_i/r_i \in K$. Let $r = r_1 + r_2$. By convexity of K, $(r_1/r)(x_1/r_1) + (r_2/r)(x_2/r_2) = (x_1 + x_2)/r \in K$. Thus, $p(x_1 + x_2)/r \leq 1$. Or by No. 2, $p(x_1 + x_2) \leq r < p(x_1) + p(x_2) + 2\varepsilon$. Since ε was arbitrary, p is subadditive.
4. Let ε be the radius of a closed sphere centered at θ and contained in K. Then for any $x \in X$, $\varepsilon x/\|x\| \in K$ and thus $p(\varepsilon x/\|x\|) \leq 1$. Hence, by No. 2, $p(x) \leq (1/\varepsilon)\|x\|$. This shows that p is continuous at θ. However, from No. 3, we have $p(x) = p(x - y + y) \leq p(x - y) + p(y)$ and $p(y) = p(y - x + x) \leq p(y - x) + p(x)$ or $-p(y - x) \leq p(x) - p(y) \leq p(x - y)$ from which continuity on X follows from continuity at θ.
5. This follows readily from No. 4. ∎

We can now prove the geometric form of the Hahn-Banach theorem.

Theorem 1. (*Mazur's Theorem, Geometric Hahn-Banach Theorem*) *Let K be a convex set having a nonempty interior in a real normed linear vector space X. Suppose V is a linear variety in X containing no interior points of K. Then there is a closed hyperplane in X containing V but containing no interior points of K; i.e., there is an element $x^* \in X^*$ and a constant c such that $\langle v, x^* \rangle = c$ for all $v \in V$ and $\langle k, x^* \rangle < c$ for all $k \in \mathring{K}$.*

Proof. By an appropriate translation we may assume that θ is an interior point of K. Let M be the subspace of X generated by V. Then V is a hyperplane in M and does not contain θ; thus there is a linear functional f on M such that $V = \{x : f(x) = 1\}$.

Let p be the Minkowski functional of K. Since V contains no interior point of K, we have $f(x) = 1 \leq p(x)$ for $x \in V$. By homogeneity, $f(\alpha x) = \alpha \leq p(\alpha x)$ for $x \in V$ and $\alpha > 0$. While for $\alpha < 0$, $f(\alpha x) \leq 0 \leq p(\alpha x)$. Thus $f(x) \leq p(x)$ for all $x \in M$. By the Hahn-Banach theorem, there is an extension F of f from M to X with $F(x) \leq p(x)$. Let $H = \{x : F(x) = 1\}$. Since $F(x) \leq p(x)$ on X and since by Lemma 1 p is continuous, F is continuous, $F(x) < 1$ for $x \in \mathring{K}$, therefore, H is the desired closed hyperplane. ∎

There are several corollaries and modifications of this important theorem, some of which are discussed in the remainder of this section.

Definition. A closed hyperplane H in a normed space X is said to be a *support* (or a *supporting hyperplane*) for the convex set K if K is contained in one of the closed half-spaces determined by H and H contains a point of $\overline{K}$.

Theorem 2. (*Support Theorem*) *If x is not an interior point of a convex set K which contains interior points, there is a closed hyperplane H containing x such that K lies on one side of H.*

As a consequence of the above theorem, it follows that, for a convex set K with interior points, a supporting hyperplane can be constructed containing any boundary point of $\overline{K}$.

Theorem 3. (*Eidelheit Separation Theorem*) *Let K_1 and K_2 be convex sets in X such that K_1 has interior points and K_2 contains no interior point of K_1. Then there is a closed hyperplane H separating K_1 and K_2; i.e., there is an $x^* \in X^*$ such that $\sup_{x \in K_1} \langle x, x^* \rangle \leq \inf_{x \in K_2} \langle x, x^* \rangle$. In other words, K_1 and K_2 lie in opposite half-spaces determined by H.*

Proof. Let $K = K_1 - K_2$; then K contains an interior point and θ is not one of them. By Theorem 2 there is an $x^* \in X^*$, $x^* \neq \theta$, such that $\langle x, x^* \rangle \leq 0$ for $x \in K$. Thus for $x_1 \in K_1$, $x_2 \in K_2$, $\langle x_1, x^* \rangle \leq \langle x_2, x^* \rangle$.

Consequently, there is a real number c such that $\sup_{K_1} \langle k_1, x^* \rangle \leq c \leq \inf_{k_2} \langle k_2, x^* \rangle$. The desired hyperplane is $H = \{x : \langle x, x^* \rangle = c\}$. ∎

Theorem 4. *If K is a closed convex set and $x \notin K$, there is a closed half-space that contains K but does not contain x.*

Proof. Let $d = \inf_{k \in K} \|x - k\|$. Then $d > 0$ since K is closed. Let S be the open sphere about x of radius $d/2$. Then apply Theorem 3 to S and K. ∎

The previous theorem can be stated in an alternative form.

Theorem 5. *If K is a closed convex set in a normed space, then K is equal to the intersection of all the closed half-spaces that contain it.*

Theorem 5 is often regarded as the geometric foundation of duality theory for convex sets. By associating closed hyperplanes (or half-spaces) with elements of X^*, the theorem expresses a convex set in X as a collection of elements in X^*.

The appeal of the above collection of theorems is that they have simple, geometrically intuitive interpretations. These apparently simple geometric facts lead to some fairly profound and useful principles of optimization.

*5.13 Duality in Minimum Norm Problems

In this section we generalize the duality principle for minimum norm problems to include the problem of finding the minimum distance from a point to a convex set. Our development of this generalization is based on the geometric notions of separating hyperplanes formulated in the last section.

The basic principle of duality is illustrated in Figure 5.4: the minimum distance from a point to a convex set K is equal to the maximum of the

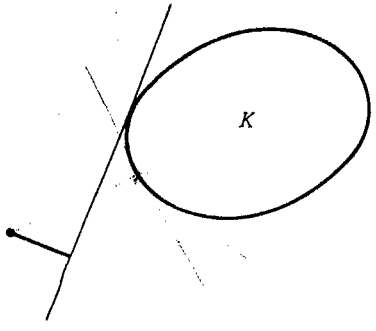

Figure 5.4 Duality

distances from the point to hyperplanes separating the point and the convex set K. We translate this simple, intuitive, geometric relation into algebraic form and show its relation to our earlier duality result. Since the results of this section are included in the more general theory of Chapter 7, and since the machinery introduced is not explicitly required for later portions of the book, the reader may wish to skip this section.

Definition. Let K be a convex set in a real normed vector space X. The functional $h(x^*) = \sup_{x \in K} \langle x, x^* \rangle$ defined on X^* is called the *support functional* of K.

In general, $h(x^*)$ may be infinite.

The support functional is illustrated in Figure 5.5. It can be interpreted geometrically as follows: Given an element $x^* \in X^*$, we consider the family of half-spaces $\{x : \langle x, x^* \rangle \leq c\}$ as the constant c varies. As c increases, these half-spaces get larger and $h(x^*)$ is defined as the infimum

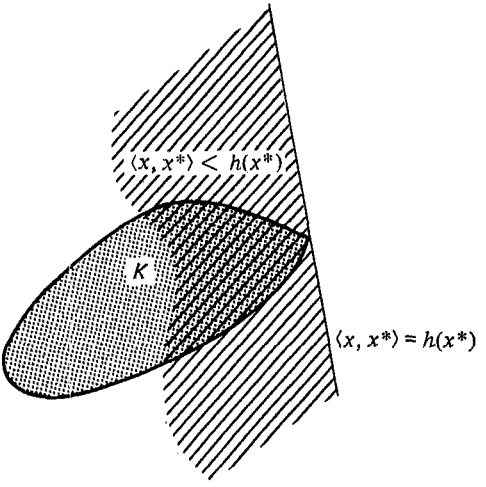

Figure 5.5 The support functional

of those constants c such that K is contained within the half-space. (The reader should verify the equivalence of this last statement with the definition above.) In view of this interpretation, it is clear that the support functional of a convex set K completely specifies the set—to within closure—since, according to Theorem 5 of the last section,

$$\overline{K} = \bigcap_{x^*} \{x : \langle x, x^* \rangle \leq h(x^*)\}.$$

The support functional has other interpretations as well, some of which are discussed in the problems. Note, in particular, that if K is the unit sphere in X, then $h(x^*)$ is simply the norm in X^*.

There is a final interpretation of the support functional that directly serves our original objective of expressing in analytical form the duality principle for the minimum norm problem illustrated in Figure 5.4. Let K be a convex set which is a finite distance from θ, let $x^* \in X^*$ have $\|x^*\| = 1$, and suppose the hyperplane $H = \{x : \langle x, x^* \rangle = h(x^*)\}$ is a support hyperplane for K separating θ from K; then the distance from θ to H is $-h(x^*)$. This interpretation is both verified and applied to the minimum norm problem in the proof of the next theorem. Figure 5.6 illustrates the result and the method of proof.

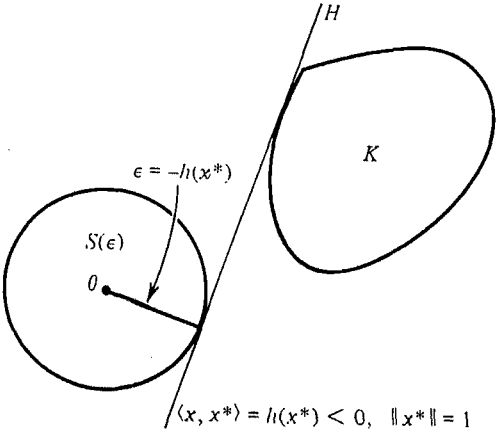

Figure 5.6 Proof of Theorem 1

Theorem 1. (*Minimum Norm Duality*) *Let x_1 be a point in a real normed vector space X and let $d > 0$ denote its distance from the convex set K having support functional h; then*

$$d = \inf_{x \in K} \|x - x_1\| = \max_{\|x^*\| \leq 1} [\langle x_1, x^* \rangle - h(x^*)]$$

where the maximum on the right is achieved by some $x_0^* \in X^*$.

If the infimum on the left is achieved by some $x_0 \in K$, then $-x_0^*$ is aligned with $x_0 - x_1$.

Proof. For simplicity we take $x_1 = 0$ since the general case can then be deduced by translation. Thus we must show that

$$d = \inf_{x \in K} \|x\| = \max_{\|x^*\| \leq 1} -h(x^*).$$

We first show that for any $x^* \in X^*$, $\|x^*\| \leq 1$, we have $d \geq -h(x^*)$. For this we may obviously limit our attention to those x^*'s which render $h(x^*)$ negative. If, however, $h(x^*)$ is negative, K is contained in the half-space $\{x : \langle x, x^* \rangle \leq h(x^*)\}$. And since $\langle \theta, x^* \rangle = 0$, this half-space does not contain θ. Therefore, if $h(x^*)$ is negative, the hyperplane $H = \{x : \langle x, x^* \rangle = h(x^*)\}$ separates K and θ.

Let $S(\varepsilon)$ be the sphere centered at θ of radius ε. For any $x^* \in X^*$, having $h(x^*) < 0$ and $\|x^*\| = 1$, let ε^* be the supremum of the ε's for which the hyperplane $\{x : \langle x, x^* \rangle = h(x^*)\}$ separates K and $S(\varepsilon)$. Obviously, we have $0 \leq \varepsilon^* \leq d$. Also $h(x^*) = \inf_{\|x\| < \varepsilon^*} \langle x, x^* \rangle = -\varepsilon^*$. Thus for every $x^* \in X^*$, $\|x^*\| \leq 1$, we have $-h(x^*) \leq d$.

On the other hand, since K contains no interior points of $S(d)$, there is a hyperplane separating $S(d)$ and K. Therefore, there is a $x_0^* \in X^*$, $\|x_0^*\| = 1$, such that $-h(x_0^*) = d$.

To prove the statement concerning alignment, suppose that $x_0 \in K$, $\|x_0\| = d$. Then $\langle x_0, x_0^* \rangle \leq h(x_0^*) = -d$ since $x_0 \in K$. However, $-\langle x_0, x_0^* \rangle \leq \|x_0^*\| \|x_0\| = d$. Thus $-\langle x_0, x_0^* \rangle = \|x_0^*\| \|x_0\|$ and $-x_0^*$ is aligned with x_0. ∎

Example 1. Suppose that $K = M$ is a subspace of the normed space X and that x is fixed in X. Theorem 1 then states that

$$\inf_{m \in M} \|x - m\| = \max_{\|x^*\| \leq 1} [\langle x, x^* \rangle - h(x^*)].$$

It is clear, however, that, corresponding to M, $h(x^*)$ is finite only for $x^* \in M^\perp$, in which case it is zero. Therefore, we obtain

$$\inf_{m \in M} \|x - m\| = \max_{\substack{\|x^*\| \leq 1 \\ x^* \in M^\perp}} \langle x, x^* \rangle$$

which is equivalent to our earlier duality result: Theorem 1, Section 5.8.

The theorem of this section is actually only an intermediate result in the development of duality that continues through the next few chapters and as such it is not of major practical importance. Nevertheless, even though this result is superseded by more general and more useful duality relations, its simplicity and geometric character make it an excellent first example of the interplay between maximization, minimization, convexity, and supporting hyperplanes.

5.14 Problems

1. Define the linear functional f on $L_2[0, 1]$ by

$$f(x) = \int_0^1 a(t) \int_0^t b(s)x(s)\, ds\, dt$$

where $a, b \in L_2[0, 1]$. Show that f is a bounded linear functional on $L_2[0, 1]$ and find an element $y \in L_2$ such that $f(x) = (x|y)$.

2. Define the Banach space c as the space of all sequences $x = \{\xi_1, \xi_2, \ldots\}$ which converge to a limit (i.e., $\lim_{k\to\infty} \xi_k$ exists), with $\|x\| = \sup_{1\le k<\infty} |\xi_k|$. Define c_0 as the space of all sequences which converge to zero (same norm as in c). Characterize the dual spaces of c_0 and c (with proofs). Warning: the dual spaces of c_0 and c are not identical.
3. Let X^* be the dual of the normed space X. Show that if X^* is separable, then X is separable.
4. Show that the normed space $C[a, b]$ is not reflexive.
5. Verify the alignment criteria given in Examples 1 and 2, Section 5.7.
6. Suppose we wish to bring a rocket car of unit mass, and subject only to the force of the rocket thrust, to rest at $x = 0$ in minimum time by proper choice of the rocket thrust program. The available thrust u is limited to $|u(t)| \le 1$ for each t. Assume that initially $x(0) = 0, \dot{x}(0) = 1$. See Figure 5.7.

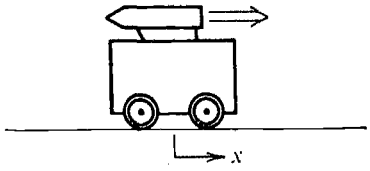

Figure 5.7 A rocket car

(a) Produce an argument that converts this problem to a minimum norm problem on a fixed interval $[0, T]$.
(b) Solve the problem.

7. The instantaneous thrust $u(t)$ produced by a rocket is a two-dimensional vector with components $u_1(t)$, $u_2(t)$. The instantaneous thrust magnitude is $|u(t)|_2 \equiv \sqrt{u_1^2(t) + u_2^2(t)}$. Consideration of the maximum magnitude over an interval of time leads to a definition of a norm as

$$\|u\| = \max_{0\le t\le 1} |u(t)|_2 .$$

Let $C_2^2[0, 1] = X$ be defined as the space of all ordered pairs $u(t) = (u_1(t), u_2(t))$ of continuous functions with norm defined as above. Show that bounded linear functionals on X can be expressed in the form

$$f(u) = \int_0^1 u_1(t)\, dv_1(t) + u_2(t)\, dv_2(t)$$

where v_1 and v_2 are functions of bounded variation and

$$\|f\| = \int_0^1 \sqrt{|dv_1|^2 + |dv_2|^2}.$$

8. Let $X = C_p^n[0, 1]$ be defined as the space of all continuous functions from $[0, 1]$ to n-dimensional space; i.e., each $x \in X$ is of the form

$$x \equiv (x_1(t), x_2(t), \ldots, x_n(t))$$

where each $x_i(t)$ is a continuous function on $[0, 1]$. The norm on X is

$$\|x\| = \sup_{0 \le t \le 1} |x(t)|_p$$

where

$$|x(t)|_p = \left\{ \sum_{i=1}^{n} x_i^p(t) \right\}^{1/p}$$

Find X^*.

9. Let x_1 and x_2 denote horizontal and vertical position components in a vertical plane. A rocket of unit mass initially at rest at the point $x_1 = x_2 = 0$ is to be propelled to the point $x_1 = x_2 = 1$ in unit time by a single jet with components of thrust u_1, u_2. Making assumptions similar to those of Example 3, Section 5.9, find the thrust program $u_1(t)$, $u_2(t)$ that accomplishes this with minimum expenditure of fuel,

$$\int_0^1 \sqrt{u_1^2(t) + u_2^2(t)}\, dt.$$

10. Let X be a normed space and M a subspace of it. Show that (within an isometric isomorphism) $M^* = X^*/M^\perp$ and $M^\perp = (X/M)^*$. (See Problem 15, Chapter 2.)
11. Let $\{x_k^*\}$ be a bounded sequence in X^* and suppose the sequence of scalars $\{\langle x, x_k^* \rangle\}$ converges for each x in a dense subset of X. Show that $\{x_k^*\}$ converges weak* to an element $x^* \in X^*$.
12. In numerical computations it is often necessary to approximate certain linear functionals by simpler ones. For instance, for $X = C[0, 1]$, we might approximate

$$L(x) = \int_0^1 x(t)\, dt$$

by a formula of the form

$$L_n(x) = \sum_{k=0}^{n} a_{nk} x\left(\frac{k}{n}\right).$$

The coefficients might be found by requiring that $L_n(p)$ be exact for all polynomials p of degree n. Following the above scheme, find $L_2(x)$ (Simpson's Rule).

13. Let $X = C[0, 1]$. Suppose there is defined the two triangular arrays of real numbers

$$\begin{array}{llll} t_{11} & & & \\ t_{21} & t_{22} & & \\ t_{31} & t_{32} & t_{33} & \\ t_{41} & \vdots & & \\ \vdots & & & \end{array} \qquad \begin{array}{lll} a_{11} & & \\ a_{21} & a_{22} & \\ a_{31} & a_{32} & a_{33} \\ a_{41} & \vdots & \\ \vdots & & \end{array}$$

and we construct the quadrature rules

$$L_n(x) = \sum_{k=1}^{n} a_{nk}\, x(t_{nk})$$

which have the property that

$$L_n(p) = \int_0^1 p(t)\, dt$$

for any polynomial p of degree $n - 1$ or less. Suppose also that

$$\sum_{k=1}^{n} |a_{nk}| < M \qquad \text{for all } n.$$

Show that for any $x \in X$

$$L_n(x) \to \int_0^1 x(t)\, dt.$$

14. Let K be a convex set in a real normed linear space X. Denote by $v(K)$ the intersection of all closed linear varieties containing K. A point x in K is said to be a relative interior point of K if x is an interior point of K, regarded as a subset of $v(K)$. See Section 2.7. Let y be a point in $v(K)$ which is not in the relative interior of K and suppose K has relative interior points. Show that there is a closed hyperplane in X containing y and having K on one side of it.
15. Let X be a real linear vector space and let $f_1, f_2, \ldots, f_n$ be linear functionals on X. Show that, for fixed α_i's, the system of equations

$$f_i(x) = \alpha_i \qquad 1 \le i \le n$$

has a solution $x \in X$ if and only if for any n numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ the relation

$$\sum_{i=1}^{n} \lambda_i f_i = 0$$

implies

$$\sum_{i=1}^{n} \lambda_i \alpha_i = 0.$$

Hint: In the space E^n consider the subspace formed by all points of the form $(f_1(x), f_2(x), \ldots, f_n(x))$ as x varies over X and apply the Hahn-Banach theorem.

16. Let $g_1, g_2, \ldots, g_n$ be linearly independent linear functionals on a vector space X. Let f be another linear functional on X such that for every $x \in X$ satisfying $g_i(x) = 0$, $i = 1, 2, \ldots, n$, we have $f(x) = 0$. Show that there are constants $\lambda_1, \lambda_2, \ldots, \lambda_n$ such that

$$f = \sum_{i=1}^{n} \lambda_i g_i.$$

17. Let X be a real linear vector space and let $f_1, f_2, \ldots, f_n$ be linear functionals on X. Show that, for fixed α_i's, the system of inequalities

$$f_i(x) \geq \alpha_i \qquad 1 \leq i \leq n$$

has a solution $x \in X$ if and only if for any n nonnegative numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ the relation

$$\sum_{i=1}^{n} \lambda_i f_i = 0$$

implies

$$\sum_{i=1}^{n} \lambda_i \alpha_i \leq 0.$$

18. Many duality properties are not completely symmetric except in reflexive spaces where $X = X^{**}$ (under the natural mapping). The theory of weak duality with a deemphasis of the role of the norm is in many respects more satisfying although of somewhat less practical importance. This problem introduces the foundation of that theory. Let X be a normed space and let X^* be its dual. Show that the weakly continuous linear functionals on X are precisely those of the form $f(x) = \langle x, x^* \rangle$ where $x^* \in X^*$ and that the weak* continuous linear functionals on X^* are precisely those of the form $g(x^*) = \langle x, x^* \rangle$ where $x \in X$. Hint: Use the result of Problem 16.
19. Show that the support functional h of a convex set is sublinear.
20. Show that the set in X^* where the support functional is finite is a convex cone.
21. Let h be the support functional of the convex set K. Find the support functional of the set $K + x_1$.

22. Prove the duality theorem of Section 5.13 for the case $x_1 \neq \theta$.
23. Let X be a real normed linear space, and let K be a convex set in X, having θ as an interior point. Let h be the support functional of K and define $K^\circ = \{x^* \in X^* : h(x^*) \leq 1\}$. Now for $x \in X$, let $p(x) = \sup_{x^* \in K^\circ} \langle x, x^* \rangle$. Show that p is equal to the Minkowski functional of K.
24. Let X, K, K°, p, h be defined as in Problem 23, with the exception that K is now an arbitrary set in X. Show that $\{x : p(x) \leq 1\}$ is equal to the closed convex hull of $K \cup \{\theta\}$.

REFERENCES

§5.2–6. The presentation of these sections closely follows that of standard works such as Yosida [157], Goffman and Pedrick [59], and Taylor [145]. These references can be consulted for additional examples of normed duals and for a proof using Zorn's lemma of the more general Hahn-Banach theorem.

§5.8–9. The duality theorems for minimum norm problems are given in the form presented here in Nirenberg [114]. This material has a parallel development known as the L-problem in the theory of moments, which was studied extensively by Akhiezer and Krein [4]. The solution of the L-problem has been applied to optimal control problems by several authors. See Krasovskii [89], Kulikowski [92], Kirillova [86], Neustadt [110], [111], and Butkovskii [26]; for an extensive bibliography, see Sarachik [137]. For additional applications of the Hahn-Banach theorem to approximation theory, see Deutsch and Maserick [39].

§5.11–12. The theory of convexity and hyperplanes has been studied extensively. For some extensions of the development given in this chapter, see Klee [87], Lorch [97], and Day [37]. For a number of counter-examples showing non-separability of disjoint convex sets when the interior point condition is not met, see Tukey [147].

§5.13. The duality theorem is proved in Nirenberg [114].

§5.14. Problems 7–9 are based on material from Neustadt [111].

6
LINEAR OPERATORS AND ADJOINTS

6.1 Introduction

A study of linear operators and adjoints is essential for a sophisticated approach to many problems of linear vector spaces. The associated concepts and notations of operator theory often streamline an otherwise cumbersome analysis by eliminating the need for carrying along complicated explicit formulas and by enhancing one's insight of the problem and its solution. This chapter contains no additional optimization principles but instead develops results of linear operator theory that make the application of optimization principles more straightforward in complicated situations. Of particular importance is the concept of the adjoint of a linear operator which, being defined in dual space, characterizes many aspects of duality theory.

Because it is difficult to obtain a simple geometric representation of an arbitrary linear operator, the material in this chapter tends to be somewhat more algebraic in character than that of other chapters. Effort is made, however, to extend some of the geometric ideas used for the study of linear functionals to general linear operators and also to interpret adjoints in terms of relations among hyperplanes.

6.2 Fundamentals

A *transformation* T is, as discussed briefly in Chapter 2, a mapping from one vector space to another. If T maps the space X into Y, we write $T: X \to Y$, and if T maps the vector $x \in X$ into the vector $y \in Y$, we write $y = T(x)$ and refer to y as the *image* of x under T. As before, we allow that a transformation may be defined only on a subset $D \subset X$, called the *domain* of T, although in most cases $D = X$. The collection of all vectors $y \in Y$ for which there is an $x \in D$ with $y = T(x)$ is called the *range* of T.

If $T: X \to Y$ and S is a given set in X, we denote by $T(S)$ the *image of* S in Y defined as the subset of Y consisting of points of the form $y = T(s)$

with $s \in S$. Similarly, given any set $P \subset Y$, we denote by $T^{-1}(P)$ the *inverse image of* P which is the set consisting of all points $x \in X$ satisfying $T(x) \in P$.

Our attention in this chapter is focused primarily on linear transformations which are alternatively referred to as *linear operators* or simply operators and are usually denoted by A, B, etc. For convenience we often omit the parentheses for a linear operator and write Ax for $A(x)$. The range of a linear operator $A : X \to Y$ is denoted $\mathscr{R}(A)$ and is obviously a subspace of Y. The set $\{x : Ax = \theta\}$ corresponding to the linear operator A is called the *nullspace* of A and denoted $\mathscr{N}(A)$. It is a subspace of X.

Of particular importance is the case in which X and Y are normed spaces and A is a continuous operator from X into Y. The following result is easily established.

Proposition 1. *A linear operator on a normed space X is continuous at every point in X if it is continuous at a single point.*

Analogous to the procedure for constructing the normed dual consisting of continuous linear functionals on a space X, it is possible to construct a normed space of continuous linear operators on X. We begin by defining the norm of a linear operator.

Definition. A linear operator A from a normed space X to a normed space Y is said to be *bounded* if there is a constant M such that $\|Ax\| \leq M\|x\|$ for all $x \in X$. The smallest such M which satisfies the above condition is denoted $\|A\|$ and called the *norm* of A.

Alternative, but equivalent, definitions of the norm are

$$\|A\| = \sup_{\|x\| \leq 1} \|Ax\|$$

$$\|A\| = \sup_{x \neq \theta} \frac{\|Ax\|}{\|x\|}.$$

We leave it to the reader to prove the following proposition.

Proposition 2. *A linear operator is bounded if and only if it is continuous.*

If addition and scalar multiplication are defined by

$$(A_1 + A_2)x = A_1x + A_2x$$

$$(\alpha A)x = \alpha(Ax)$$

the linear operators from X to Y form a linear vector space. If X and Y are normed spaces, the subspace of continuous linear operators can be identified and this becomes a normed space when the norm of an operator

is defined according to the last definition. (The reader can easily verify that the requirements for a norm are satisfied.)

Definition. The normed space of all bounded linear operators from the normed space X into the normed space Y is denoted $B(X, Y)$.

We note the following result which generalizes Theorem 1, Section 5.2. The proof requires only slight modification of the proof in Section 5.2 and is omitted here.

Theorem 1. *Let X and Y be normed spaces with Y complete. Then the space $B(X, Y)$ is complete.*

In general the space $B(X, Y)$, although of interest by its own right, does not play nearly as dominant a role in our theory as that of the normed dual of X. Nevertheless, certain of its elementary properties and the definition itself are often convenient. For instance, we write $A \in B(X, Y)$ for, "let A be a continuous linear operator from the normed space X to the normed space Y."

Finally, before turning to some examples, we observe that the spaces of linear operators have a structure not present in an arbitrary vector space in that it is possible to define products of operators. Thus, if $S : X \to Y$, $T : Y \to Z$, we define the operator $TS : X \to Z$ by the equation $(TS)(x) = T(Sx)$ for all $x \in X$. For bounded operators we have the following useful result.

Proposition 3. *Let X, Y, Z be normed spaces and suppose $S \in B(X, Y)$, $T \in B(Y, Z)$. Then $\|TS\| \leq \|T\|\|S\|$.*

Proof. $\|TSx\| \leq \|T\|\|Sx\| \leq \|T\|\|S\|\|x\|$ for all $x \in X$. ∎

Example 1. Let $X = C[0, 1]$ and define the operator $A : X \to X$ by $Ax = \int_0^1 K(s, t)x(t)\,dt$ where the function K is continuous on the unit square $0 \leq s \leq 1$, $0 \leq t \leq 1$. The operator A is clearly linear. We compute $\|A\|$. We have

$$\begin{aligned}\|Ax\| &= \max_{0 \leq s \leq 1} \left| \int_0^1 K(s, t)x(t)\,dt \right| \\ &\leq \max_{0 \leq s \leq 1} \left\{ \int_0^1 |K(s, t)|\,dt \right\} \max_{0 \leq t \leq 1} |x(t)| \\ &= \max_{0 \leq s \leq 1} \int_0^1 |K(s, t)|\,dt \cdot \|x\|.\end{aligned}$$

Therefore,

$$\|A\| \leq \max_{0 \leq s \leq 1} \int_0^1 |K(s, t)|\,dt.$$

We can show that the quantity on the right-hand side is actually the norm of A. Let s_0 be the point at which the continuous function

$$\int_0^1 |K(s, t)|\, dt$$

achieves its maximum. Given $\varepsilon > 0$ let p be a polynomial which approximates $K(s_0, \cdot)$ in the sense that

$$\max_{0 \le t \le 1} |K(s_0, t) - p(t)| < \varepsilon$$

and let x be a function in $C[0, 1)$ with $\|x\| \le 1$ which approximates the discontinuous function sgn $p(t)$ in the sense that

$$\left| \int_0^1 p(t)x(t)\, dt - \int_0^1 |p(t)|\, dt \right| < \varepsilon.$$

This last approximation is easily constructed since p has only a finite number of sign changes.

For this x we have

$$\begin{aligned}
\left| \int_0^1 K(s_0, t)x(t)\, dt \right| &\ge \left| \int_0^1 p(t)x(t)\, dt \right| - \left| \int_0^1 [K(s_0, t) - p(t)]x(t)\, dt \right| \\
&\ge \left| \int_0^1 p(t)x(t)\, dt \right| - \varepsilon \ge \int_0^1 |p(t)|\, dt - 2\varepsilon \\
&\ge \int_0^1 |K(s_0, t)|\, dt - \left| \int_0^1 [|K(s_0, t)| - |p(t)|]\, dt \right| - 2\varepsilon \\
&\ge \int_0^1 |K(s_0, t)|\, dt - 3\varepsilon.
\end{aligned}$$

Thus, since $\|x\| \le 1$,

$$\|A\| \ge \int_0^1 |K(s_0, t)|\, dt - 3\varepsilon.$$

But since ε was arbitrary, and since the reverse inequality was established above, we have

$$\|A\| = \max_{0 \le s \le 1} \int_0^1 |K(s, t)|\, dt.$$

Example 2. Let $X = E^n$ and let $A : X \to X$. Then A is a matrix acting on the components of x. We have $\|Ax\|^2 = (x \mid A'Ax)$ where A' is the transpose of the matrix A. Denoting $A'A$ by Q, determination of $\|A\|$ is equivalent to maximizing $(x \mid Qx)$ subject to $\|x\|^2 \le 1$. This is a finite-dimensional optimization problem. Since Q is symmetric and positive semidefinite, it

has nonnegative eigenvalues and the solution of the optimization problem is given by x equal to the eigenvector of Q corresponding to the largest eigenvalue.

We conclude that $\|A\| = \sqrt{\lambda_{\max}}$.

Example 3. The operator $Ax = d/dt\, x(t)$, defined on the subspace M of $C[0, 1]$ consisting of all continuously differentiable functions, has range $C[0, 1]$. A is not bounded, however, since elements of arbitrarily small norm can produce elements of large norm when differentiated. On the other hand, if A is regarded as having domain $D[0, 1]$ and range $C[0, 1]$, it is bounded with $\|A\| = 1$.

INVERSE OPERATORS

6.3 Linearity of Inverses

Let $A : X \to Y$ be a linear operator between two linear spaces X and Y. Corresponding to A we consider the equation $Ax = y$. For a given $y \in Y$ this equation may:

1. have a unique solution $x \in X$,
2. have no solution,
3. have more than one solution.

Many optimization problems can be regarded as arising from cases 2 or 3; these are discussed in Section 6.9. Condition 1 holds for every $y \in Y$ if and only if the mapping A from X to Y is one-to-one and has range equal to Y, in which case the operator A has an *inverse* A^{-1} such that if $Ax = y$, then $A^{-1}(y) = x$.

Proposition 1. *If a linear operator $A : X \to Y$ has an inverse, the inverse A^{-1} is linear.*

Proof. Suppose $A^{-1}(y_1) = x_1$, $A^{-1}(y_2) = x_2$, then

$$A(x_1) = y_1, \qquad A(x_2) = y_2,$$

and the linearity of A implies that $A(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 y_1 + \alpha_2 y_2$. Thus $A^{-1}(\alpha_1 y_1 + \alpha_2 y_2) = \alpha_1 A^{-1}(y_1) + \alpha_2 A^{-1}(y_2)$. ∎

The solution of linear equations and the determination of inverse operators are, of course, important areas of pure and applied mathematics. For optimization theory, however, we are not so much interested in solving equations as formulating the equations appropriate for characterizing an optimal vector. Once the equations are formulated, we may rely on standard techniques for their solution. There are important exceptions to this point

of view, however, since optimization theory often provides effective procedures for solving equations. Furthermore, a problem can never really be regarded as resolved until an efficient computational method of solution is derived. Nevertheless, our primary interest in linear operators is their role in optimization problems. We do not develop an extensive theory of linear equations but are content with establishing the existence of a solution.

6.4 The Banach Inverse Theorem

Given a continuous linear operator A from a normed space X onto a normed space Y and assuming that A has an inverse A^{-1}, it follows that A^{-1} is linear but not necessarily continuous. If, however, X and Y are Banach spaces, A^{-1} must be continuous if it exists. This result, known as the Banach inverse theorem, is one of the analytical cornerstones of functional analysis. Many important, deep, and sometimes surprising results follow from it. We make application of the result in Section 6.6 and again in Chapter 8 in connection with Lagrange multipliers. Other applications to problems of mathematical analysis are discussed in the problems at the end of this chapter.

This section is devoted to establishing this one result. Although the proof is no more difficult at each step than that of most theorems in this book, it involves a number of steps. Therefore, since it plays only a supporting role in the optimization theory, the reader may wish to simply scan the proof and proceed to the next section.

We begin by establishing the following lemma which itself is an important and celebrated tool of analysis.

Lemma 1. (*Baire*) *A Banach space X is not the union of countably many nowhere dense sets in X.*

Proof. Suppose that $\{E_n\}$ is a sequence of nowhere dense sets and let F_n denote the closure of E_n. Then F_n contains no sphere in X. It follows that each of the sets $\tilde{F}_n$ is open and dense in X.

Let $S(x_1, r_1)$ be a sphere in $\tilde{F}_1$ with center at x_1 and radius r_1. Let $S(x_2, r_2)$ be a sphere in $\tilde{F}_2 \cap S(x_1, r_1/2)$. (Such a sphere exists since $\tilde{F}_2$ is open and dense.) Proceeding inductively, let $S(x_n, r_n)$ be a sphere in $S(x_{n-1}, r_{n-1}/2) \cap \tilde{F}_n$.

The sequence $\{x_n\}$ so defined is clearly a Cauchy sequence and, thus, by the completeness of X, there is a limit x; $x_n \to x$. This vector x lies in each of the $S(x_n, r_n)$, because, indeed, $x_{n+k} \in S(x_n, r_n/2)$ for $k \geq 1$. Hence x lies in each $\tilde{F}_n$. Therefore, $x \in \bigcap_n \tilde{F}_n$.

It follows that the union of the original collection of sets $\{E_n\}$ is not X since

$$x \in \bigcap_n \tilde{F}_n = \left[\bigcup_n \widetilde{F_n}\right] \subset \left[\bigcup_n \widetilde{E_n}\right]. \blacksquare$$

Theorem 1. (*Banach Inverse Theorem*) *Let A be a continuous linear operator from a Banach space X onto a Banach space Y and suppose that the inverse operator A^{-1} exists. Then A^{-1} is continuous.*

Proof. In view of the linearity of A and therefore of A^{-1}, it is only necessary to show that A^{-1} is bounded. For this it is only necessary to show that the image $A(S)$ in Y of any sphere S centered at the origin in X contains a sphere P centered at the origin in Y, because then the inverse image of P is contained in S. The proof amounts to establishing the existence of a sphere in $A(S)$.

Given a sphere S, for any $x \in X$ there is an integer n such that $x/n \in S$ and hence $A(x/n) \in A(S)$ or, equivalently, $A(x) \in nA(S)$. Since A maps X onto Y, it follows that

$$Y = \bigcup_{n=1}^{\infty} nA(S).$$

According to Baire's lemma, Y cannot be the union of countably many nowhere dense sets and, hence, there is an n such that the closure of $nA(S)$ contains a sphere. It follows that $\overline{A(S)}$ contains a sphere whose center y may be taken to be in $A(S)$. Let this sphere $N(y, r)$ have radius r, and let $y = A(x)$. Now as y' varies over $N(y, r)$, the points $y' - y$ cover the sphere $N(\theta, r)$ and the points of a dense subset of these are of the form $A(x' - x)$ where $A(x') = y'$, $x' \in S$. Since $x', x \in S$, it follows that $x' - x \in 2S$. Hence, the closure of $A(2S)$ contains $N(\theta, r)$ (and by linearity $\overline{A(S)}$ contains $N(\theta, r/2)$).

We have shown that the closure of the image of a sphere centered at the origin contains such a sphere in Y, but it remains to be shown that the image itself, rather than its closure, contains a sphere. For any $\varepsilon > 0$, let $S(\varepsilon)$ and $P(\varepsilon)$ be the spheres in X, Y, respectively, of radii ε centered at the origins. Let $\varepsilon_0 > 0$ be arbitrary and let $\eta_0 > 0$ be chosen so that $P(\eta_0)$ is a sphere contained in the closure of the image of $S(\varepsilon_0)$. Let y be an arbitrary point in $P(\eta_0)$. We show that there is an $x \in S(2\varepsilon_0)$ such that $Ax = y$ so that the image of the sphere of radius $2\varepsilon_0$ contains the sphere $P(\eta_0)$.

Let $\{\varepsilon_i\}$ be a sequence of positive numbers such that $\sum_{i=1}^{\infty} \varepsilon_i < \varepsilon_0$. Then there is a sequence $\{\eta_i\}$, with $\eta_i > 0$ and $\eta_i \to 0$, such that

$$P(\eta_i) \subset \overline{A[S(\varepsilon_i)]}$$

Since $A(S(\varepsilon_0))$ is dense in $P(\varepsilon_0)$, there is an $x_0 \in S(\varepsilon_0)$ such that $y - Ax_0 \in P(\eta_1)$. It follows that there is an $x_1 \in S(\varepsilon_1)$ with $y - Ax_0 - Ax_1 \in P(\eta_2)$. Proceeding inductively, a sequence $\{x_n\}$ is defined with $x_n \in S(\varepsilon_n)$ and $y - A(\sum_{i=0}^{n} x_i) \in P(\eta_{n+1})$. Let $z_n = x_0 + x_1 + \cdots + x_n$. Then evidently $\{z_n\}$ is a Cauchy sequence since for $m > n$, $\|z_m - z_n\| = \|x_{n-1} + x_{n-2} + \cdots + x_m\| < \varepsilon_{n+1} + \varepsilon_{n+2} + \cdots + \varepsilon_m$. Thus there is an $x \in X$ such that $z_n \to x$. Furthermore, $\|x\| < \varepsilon_0 + \varepsilon_1 + \cdots + \varepsilon_n + \cdots < 2\varepsilon_0$; so $x \in S(2\varepsilon_0)$. Since A is continuous, $Az_n \to Ax$, but since $\|y - Az_n\| < \eta_{n+1} \to 0$, $Az_n \to y$. Therefore, $Ax = y$. ∎

ADJOINTS

6.5 Definition and Examples

The constraints imposed in many optimization problems by differential equations, matrix equations, etc., can be described by linear operators. The resolution of these problems almost invariably calls for consideration of an associated operator: the adjoint. The reason for this is that adjoints provide a convenient mechanism for describing the orthogonality and duality relations which permeate nearly every optimization analysis.

Definition. Let X and Y be normed spaces and let $A \in B(X, Y)$. The *adjoint operator* $A^*: Y^* \to X^*$ is defined by the equation

$$\langle x, A^*y^* \rangle = \langle Ax, y^* \rangle.$$

This important definition requires a bit of explanation and justification. Given a fixed $y^* \in Y^*$, the quantity $\langle Ax, y^* \rangle$ is a scalar for each $x \in X$ and is therefore a functional on X. Furthermore, by the linearity of y^* and A, it follows that this functional is linear. Finally, since

$$|\langle Ax, y^* \rangle| \leq \|y^*\| \|Ax\| \leq \|y^*\| \|A\| \|x\|,$$

it follows that this functional is bounded and is thus an element x^* of X^*. We then define $A^*y^* = x^*$. The adjoint is obviously unique and the reader can verify that it is linear. It is important to remember, as illustrated in Figure 6.1, that $A^*: Y^* \to X^*$.

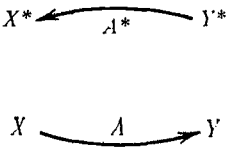

Figure 6.1 An operator and its adjoint

In terms of operator, rather than bracket, notation the definition of the adjoint satisfies the equation

$$y^*(Ax) = (A^*y^*)(x)$$

for each $x \in X$. Thus we may write

$$y^*A = A^*y^*$$

where the left side denotes the functional on X which is the composition of the operators A and y^* and the right side is the functional obtained by operating on y^* by A^*.

Theorem 1. *The adjoint operator A^* of the linear operator $A \in B(X, Y)$ is linear and bounded with $\|A^*\| = \|A\|$.*

Proof. The proof of linearity is elementary and left to the reader. From the inequalities

$$|\langle x, A^*y^*\rangle| = |\langle Ax, y^*\rangle| \leq \|y^*\|\|Ax\| \leq \|y^*\|\|A\|\|x\|$$

it follows that

$$\|A^*y^*\| \leq \|A\|\|y^*\|$$

which implies that

$$\|A^*\| \leq \|A\|.$$

Now let x_0 be any nonzero element of X. According to Corollary 2 of the Hahn-Banach theorem, there exists an element $y_0^* \in Y^*$, $\|y_0^*\| = 1$, such that $\langle Ax_0, y_0^*\rangle = \|Ax_0\|$. Therefore,

$$\|Ax_0\| = |\langle x_0, A^*y_0^*\rangle| \leq \|A^*y_0^*\|\|x_0\| \leq \|A^*\|\|x_0\|$$

from which we conclude that

$$\|A\| \leq \|A^*\|$$

It now follows that $\|A^*\| = \|A\|$. ∎

In addition to the above result, adjoints enjoy the following algebraic relations which follow easily from the basic definition.

Proposition 1. *Adjoints satisfy the following properties:*

1. *If I is the identity operator on a normed space X, then $I^* = I$.*
2. *If $A_1, A_2 \in B(X, Y)$, then $(A_1 + A_2)^* = A_1^* + A_2^*$.*
3. *If $A \in B(X, Y)$ and α is a real scalar, then $(\alpha A)^* = \alpha A^*$.*
4. *If $A_1 \in B(X, Y)$, $A_2 \in B(Y, Z)$, then $(A_2 A_1)^* = A_1^* A_2^*$.*
5. *If $A \in B(X, Y)$ and A has a bounded inverse, then $(A^{-1})^* = (A^*)^{-1}$.*

Proof. Properties 1–4 are trivial. To prove property 5, let $A \in B(X, Y)$ have a bounded inverse A^{-1}. To show that A^* has an inverse, we must show that it is one-to-one and onto. Let $y_1^* \neq y_2^* \in Y^*$, then

$$\langle x, A^* y_1^* \rangle - \langle x, A^* y_2^* \rangle = \langle Ax, y_1^* - y_2^* \rangle \neq 0$$

for some $x \in X$. Thus, $A^* y_1^* \neq A^* y_2^*$ and A^* is one-to-one. Now for any $x^* \in X^*$ and any $x \in X$, $Ax = y$, we have

$$\begin{aligned} \langle x, x^* \rangle = \langle A^{-1} y, x^* \rangle &= \langle y, (A^{-1})^* x^* \rangle \\ &= \langle Ax, (A^{-1})^* x^* \rangle = \langle x, A^* (A^{-1})^* x^* \rangle \end{aligned}$$

which shows that x^* is in $\mathscr{R}(A^*)$ and also that $(A^*)^{-1} = (A^{-1})^*$. ∎

An important special case is that of a linear operator $A : H \to G$ where H and G are Hilbert spaces. If H and G are real, then they are their own duals in the sense of Section 5.3, and the operator A^* can be regarded as mapping G into H. In this case the adjoint relation becomes $(Ax \mid y) = (x \mid A^* y)$. If the spaces are complex, the adjoint, as defined earlier, does not satisfy this relation and it is convenient and customary to redefine the Hilbert space adjoint directly by the relation $(Ax \mid y) = (x \mid A^* y)$. In our study, however, we restrict our attention to real spaces so that difficulties of this nature can be ignored.

Note that in Hilbert space we have the additional property: $A^{**} = A$.

Finally, we note the following two definitions.

Definition. A bounded linear operator A mapping a real Hilbert space into itself is said to be *self-adjoint* if $A^* = A$.

Definition. A self-adjoint linear operator A on a Hilbert space H is said to be *positive semidefinite* if $(x \mid Ax) \geq 0$ for all $x \in H$.

Example 1. Let $X = Y = E^n$. Then $A : X \to X$ is represented by an $n \times n$ matrix. Thus the i-th component of Ax is

$$(Ax)_i = \sum_{j=1}^{n} a_{ij} x_j .$$

We compute A^*. For $y \in Y$ we have

$$(Ax \mid y) = \sum_{i=1}^{n} \sum_{j=1}^{n} y_i a_{ij} x_j = \sum_{j=1}^{n} x_j \sum_{i=1}^{n} a_{ij} y_i = (x \mid A^* y)$$

where A^* is the matrix with elements $a_{ij}^* = a_{ji}$. Thus A^* is the transpose of A.

Example 2. Let $X = Y = L_2[0, 1]$ and define

$$Ax = \int_0^1 K(t, s) x(s)\, ds, \qquad t \in [0, 1]$$

where

$$\int_0^1 \int_0^1 |K(t, s)|^2 \, ds \, dt < \infty.$$

Then

$$(Ax \mid y) = \int_0^1 y(t) \int_0^1 K(t, s)x(s) \, ds \, dt$$

$$= \int_0^1 x(s) \int_0^1 K(t, s)y(t) \, dt \, ds.$$

Or, by interchanging the roles of s and t,

$$(Ax \mid y) = \int_0^1 x(t) \int_0^1 K(s, t)y(s) \, ds \, dt = (x \mid A^*y)$$

where

$$A^*y = \int_0^1 K(s, t)y(s) \, ds.$$

Therefore, the adjoint of A is obtained by interchanging s and t in K.

Example 3. Again let $X = Y = L_2[0, 1]$ and define

$$Ax = \int_0^t K(t, s)x(s) \, ds, \qquad t \in [0, 1],$$

with

$$\int_0^1 \int_0^1 |K(t, s)|^2 \, dt \, ds < \infty.$$

Then

$$(Ax \mid y) = \int_0^1 y(t) \int_0^t K(t, s)x(s) \, ds \, dt$$

$$= \int_0^1 \int_0^t y(t)K(t, s)x(s) \, ds \, dt.$$

The double integration represents integration over the triangular region shown in Figure 6.2*a*, integrating vertically and then horizontally. Alternatively, the integration may be performed in the reverse order as in Figure 6.2*b*, leading to

$$(Ax \mid y) = \int_0^1 \int_s^1 y(t)K(t, s)x(s) \, dt \, ds$$

$$= \int_0^1 x(s) \left(\int_s^1 K(t, s)y(t) \, dt \right) ds.$$

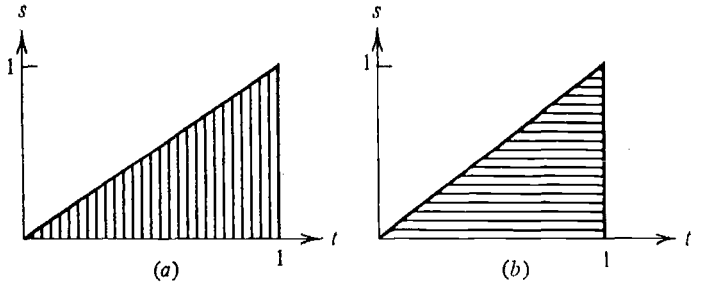

Figure 6.2 Region of integration

Or, interchanging the roles of t and s,

$$(Ax \mid y) = \int_0^1 x(t)\left(\int_t^1 K(s, t)y(s)\, ds\right) dt = (x \mid A^*y)$$

where

$$A^*y = \int_t^1 K(s, t)y(s)\, ds.$$

This example comes up frequently in the study of dynamic systems.

Example 4. Let $X = C[0, 1]$, $Y = E^n$ and define $A : X \to Y$ by the equation

$$Ax = (x(t_1), x(t_2), \ldots, x(t_n))$$

where $0 \leq t_1 < t_2 < t_2 < \cdots < t_n \leq 1$ are fixed. It is easily verified that A is continuous and linear. Let $y^* = (y_1, y_2, \ldots, y_n)$ be a linear functional on E^n. Then

$$\langle Ax, y^* \rangle = \sum_{i=1}^{n} y_i x(t_i) = \int_0^1 x(t)\, dv(t) = \langle x, A^*y^* \rangle$$

where $v(t)$ is constant except at the points t_i where it has a jump of magnitude y_i, as illustrated in Figure 6.3. Thus $A^* : E^n \to NBV[0, 1]$ is defined by $A^*y^* = v$.

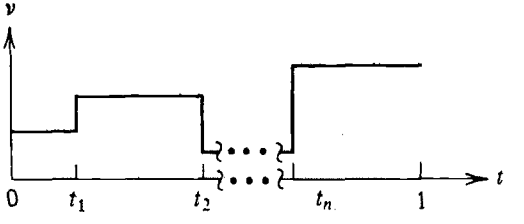

Figure 6.3 The function v

6.6 Relations Between Range and Nullspace

Adjoints are extremely useful in our recurring task of translating between the geometric properties and the algebraic description of a given problem. The following theorem and others similar to it are of particular interest.

Theorem 1. *Let X and Y be normed spaces and let $A \in B(X, Y)$. Then*

$$[\mathscr{R}(A)]^{\perp} = \mathscr{N}(A^*).$$

Proof. Let $y^* \in \mathscr{N}(A^*)$ and $y \in \mathscr{R}(A)$. Then $y = Ax$ for some $x \in X$. The calculation $\langle y, y^* \rangle = \langle Ax, y^* \rangle = \langle x, A^*y^* \rangle = 0$ shows that $\mathscr{N}(A^*) \subset [\mathscr{R}(A)]^{\perp}$.

Now assume $y^* \in [\mathscr{R}(A)]^{\perp}$. Then for every $x \in X$, $\langle Ax, y^* \rangle = 0$. This implies $\langle x, A^*y^* \rangle = 0$ and hence that $[\mathscr{R}(A)]^{\perp} \subset \mathscr{N}(A^*)$. ∎

Example 1. Let us consider the finite-dimensional version of Theorem 1. Let A be a matrix; $A: E^n \to E^m$. A consists of n column vectors a_i, $i = 1, 2, \ldots, n$, and $\mathscr{R}(A)$ is the subspace of E^m spanned by these vectors. $[\mathscr{R}(A)]^{\perp}$ consists of those vectors in E^m that are orthogonal to each a_i.

On the other hand, the matrix A^* (which is just the transpose of A) has the a_i's as its rows; hence the vectors in E^m orthogonal to a_i's comprise the nullspace of A^*. Therefore, both $[\mathscr{R}(A)]^{\perp}$ and $\mathscr{N}(A^*)$ consist of all vectors orthogonal to each a_i.

Our next theorem is a dual to Theorem 1. It should be noted, however, that the additional hypothesis that $\mathscr{R}(A)$ be closed, is required. Moreover, the dual theorem is much deeper than Theorem 1, since the proof requires both the Banach inverse theorem and the Hahn-Banach theorem.

Lemma 1. *Let X and Y be Banach spaces and let $A \in B(X, Y)$. Assume that $\mathscr{R}(A)$ is closed. Then there is a constant K such that for each $y \in \mathscr{R}(A)$ there is an x satisfying $Ax = y$ and $\|x\| \leq K\|y\|$.*

Proof. Let $N = \mathscr{N}(A)$ and consider the space X/N consisting of equivalence classes $[x]$ modulo N. Define $\bar{A}: X/N \to \mathscr{R}(A)$ by $\bar{A}[x] = Ax$. It is easily verified that $\bar{A}$ is one-to-one, onto, linear, and bounded. Since $\mathscr{R}(A)$ closed implies that $\mathscr{R}(A)$ is a Banach space, it follows from the Banach inverse theorem that $\bar{A}$ has a continuous inverse. Hence, given $y \in \mathscr{R}(A)$, there is $[x] \in X/N$ with $\|[x]\| \leq \|\bar{A}^{-1}\|\|y\|$. Take $x \in [x]$ with $\|x\| \leq 2\|[x]\|$ and then $K = 2\|\bar{A}^{-1}\|$ satisfies the conditions stated in the lemma. ∎

Now we give the dual to Theorem 1.

Theorem 2. *Let X and Y be Banach spaces and let $A \in B(X, Y)$. Let $\mathscr{R}(A)$ be closed. Then*

$$\mathscr{R}(A^*) = [\mathscr{N}(A)]^\perp.$$

Proof. Let $x^* \in \mathscr{R}(A^*)$. Then $x^* = A^*y^*$ for some $y^* \in Y^*$. For any $x \in \mathscr{N}(A)$, we have

$$\langle x, x^* \rangle = \langle x, A^*y^* \rangle = \langle Ax, y^* \rangle = 0.$$

Thus $x^* \in [\mathscr{N}(A)]^\perp$ and it follows that $\mathscr{R}(A^*) \subset [\mathscr{N}(A)]^\perp$.

Now assume that $x^* \in [\mathscr{N}(A)]^\perp$. For $y \in \mathscr{R}(A)$ and each x satisfying $Ax = y$, the functional $\langle x, x^* \rangle$ has the same value. Hence, define $f(y) = \langle x, x^* \rangle$ on $\mathscr{R}(A)$. Let K be defined as in the lemma. Then for each $y \in \mathscr{R}(A)$ there is an x with $\|x\| \leq K\|y\|$, $Ax = y$. Therefore, $|f(y)| \leq K\|x^*\|\|y\|$ and thus f is a bounded linear functional on $\mathscr{R}(A)$. Extend f by the Hahn-Banach theorem to a functional $y^* \in Y^*$. Then from

$$\langle x, A^*y^* \rangle = \langle Ax, y^* \rangle = \langle x, x^* \rangle,$$

it follows that $A^*y^* = x^*$ and thus $\mathscr{R}(A^*) \supset [\mathscr{N}(A)]^\perp$. ∎

In many applications the range of the underlying operator is finite dimensional, and hence satisfies the closure requirement. In other problems, however, this requirement is not satisfied and this generally leads to severe analytical difficulties. We give an example of an operator whose range is not closed.

Example 2. Let $X = Y = l_1$ with $A: X \to Y$ defined by

$$A\{\xi_1, \xi_2, \ldots, \xi_n, \ldots\} = \left\{\xi_1, \frac{1}{2}\xi_2, \frac{1}{3}\xi_3, \cdots, \frac{1}{n}\xi_n, \cdots\right\}.$$

Then $\mathscr{R}(A)$ contains all finitely nonzero sequences and thus $\overline{\mathscr{R}(A)} = Y$. However,

$$y = \left\{1, \frac{1}{2^2}, \frac{1}{3^2}, \cdots, \frac{1}{n^2}, \cdots\right\} \notin \mathscr{R}(A)$$

and thus $\mathscr{R}(A)$ is not closed.

In Hilbert space there are several additional useful relations between range and nullspace similar to those which hold in general normed space. These additional properties are a consequence of the fact that in Hilbert space an operator and its adjoint are defined on the same space.

Theorem 3. *Let A be a bounded linear operator acting between two real Hilbert spaces. Then*

1. $[\mathscr{R}(A)]^{\perp} = \mathscr{N}(A^*)$.
2. $\overline{\mathscr{R}(A)} = [\mathscr{N}(A^*)]^{\perp}$.
3. $[\mathscr{R}(A^*)]^{\perp} = \mathscr{N}(A)$.
4. $\overline{\mathscr{R}(A^*)} = [\mathscr{N}(A)]^{\perp}$.

Proof. Part 1 is just Theorem 1. To prove part 2, take the orthogonal complement of both sides of 1 obtaining $[\mathscr{R}(A)]^{\perp\perp} = [\mathscr{N}(A^*)]^{\perp}$. Since $\mathscr{R}(A)$ is a subspace, the result follows. Parts 3 and 4 are obtained from 1 and 2 by use of the relation $A^{**} = A$. ∎

Example 3. Let $X = Y = l_2$. For $x = \{\xi_1, \xi_2 \ldots,\}$, define $Ax = \{0, \xi_1, \xi_2, \ldots\}$. A is a shift operator (sometimes referred to as the creation operator because a new component is created). The adjoint of A is easily computed to be the operator taking $y = \{\eta_1, \eta_2, \ldots\}$ into $A^*y = \{\eta_2, \eta_3, \ldots\}$, which is a shift in the other direction (referred to as the destruction operator). It is clear that $[\mathscr{R}(A)]^{\perp}$ consists of all those vectors in l_2 that are zero except possibly in their first component; this subspace is identical with $\mathscr{N}(A^*)$.

6.7 Duality Relations for Convex Cones

The fundamental algebraic relations between nullspace and range for an operator and its adjoint derived in Section 6.6 have generalizations which often play a role in the analysis of problems described by linear inequalities analogous to the role of the earlier results to problems described by linear equalities.

Definition. Given a set S in a normed space X, the set $S^{\oplus} = \{x^* \in X^* : \langle x, x^* \rangle \geq 0 \text{ for all } x \in S\}$ is called the *positive conjugate cone of* S. Likewise the set $S^{\ominus} = \{x^* \in X^* : \langle x, x^* \rangle \leq 0 \text{ for all } x \in S\}$ is called the *negative conjugate cone of* S.

It is a simple matter to verify that $S^{\oplus}$ and $S^{\ominus}$ are in fact convex cones. They are nonempty since they always contain the zero functional. If S is a subspace of X, then obviously $S^{\oplus} = S^{\ominus} = S^{\perp}$; hence, the conjugate cones can be regarded as generalizations of the orthogonal complement of a set. The definition is illustrated in Figure 6.4 for the Hilbert space situation where $S^{\oplus}$ and $S^{\ominus}$ can be regarded as subsets of X. The basic properties

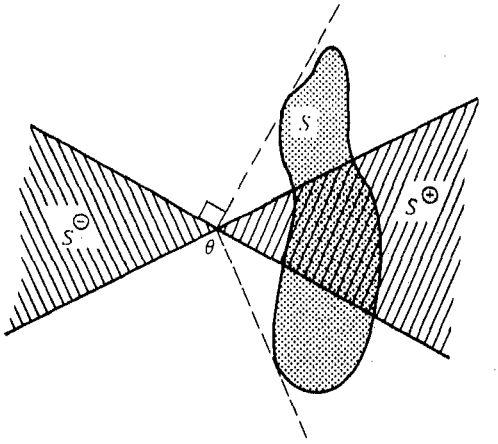

Figure 6.4 A set and its conjugate cones

of the operation of taking conjugate cones are given in the following proposition.

Proposition 1. *Let S and T be sets in a normed space X. Then*

1. $S^{\oplus}$ *is a closed convex cone in* X^*.
2. *If* $S \subset T$, *then* $T^{\oplus} \subset S^{\oplus}$.

In the general case the conjugate cone can be interpreted as a collection of half-spaces. If $x^* \in S^{\oplus}$, then clearly $\inf_{x \in S} \langle x, x^* \rangle \geq 0$ and hence the hyperplane $\{x : \langle x, x^* \rangle = 0\}$ has S in its positive half-space. Conversely, if x^* determines a hyperplane having S in its positive half-space, it is a member of $S^{\oplus}$. Therefore, $S^{\oplus}$ consists of all x^* which contain S in their positive half-spaces.

The following theorem generalizes Theorem 1 of Section 6.6.

Theorem 1. *Let X and Y be normed linear spaces and let* $A \in B(X, Y)$. *Let S be a subset of X. Then*

$$[A(S)]^{\oplus} = A^{*-1}(S^{\oplus})$$

(*where the inverse denotes the inverse image of* $S^{\oplus}$).

Proof. Assume $y^* \in [A(S)]^{\oplus}$ and $s \in S$. Then $\langle As, y^* \rangle \geq 0$ and hence $\langle s, A^* y^* \rangle \geq 0$. Thus, since s is arbitrary in S, $y^* \in A^{*-1}(S^{\oplus})$. The argument is reversible. ∎

Note that by putting $S = X$, $S^{\oplus} = \{\theta\}$, the above result reduces to $[\mathscr{R}(A)]^{\perp} = \mathscr{N}(A^*)$.

*6.8 Geometric Interpretation of Adjoints

It is somewhat difficult to obtain a clear simple visualization of the relation between an operator and its adjoint since if $A : X \to Y$, $A^* : Y^* \to X^*$, four spaces and two operators are involved. However, in view of the unique correspondence between hyperplanes not containing the origin in a space and nonzero elements of its dual, the adjoint A^* can be regarded as mapping hyperplanes in Y into hyperplanes in X. This observation can be used to consolidate the adjoint relations into two spaces rather than four. We limit our discussion here to invertible operators between Banach spaces. The arguments can be extended to the more general case, but the picture becomes somewhat more complex.

Let us fix our attention on a given hyperplane $H \subset X$ having $\theta \notin H$. The operator A maps this hyperplane point by point into a subset L of Y. It follows from the linearity of A that L is a linear variety, and since A is assumed to be invertible, it follows that L is in fact a hyperplane in Y not containing $\theta \in Y$. Therefore, A maps the hyperplane H point by point into a hyperplane L.

The hyperplanes H and L define unique elements $x_1^* \in X^*$ and $y_1^* \in Y^*$ through the relations $H = \{x : \langle x, x_1^* \rangle = 1\}$, $L = \{y : \langle y, y_1^* \rangle = 1\}$. The adjoint operator A^* can then be applied to y_1^* to produce an x_1^* or, equivalently, A^* maps L into a hyperplane in X. In fact, A^* maps L back to H. For if $A^* y_1^* = x_1^*$, it follows directly from the definition of adjoints that $\{x : \langle x, x_1^* \rangle = 1\} = \{x : \langle x, A^* y_1^* \rangle = 1\} = \{x : \langle Ax, y_1^* \rangle = 1\} = H$. Therefore, A^* maps the hyperplane L, as a unit, back to the hyperplane H. This interpretation is illustrated in Figure 6.5 where the dotted line arrows symbolize elements of a dual space.

Another geometric interpretation is discussed in Problem 13.

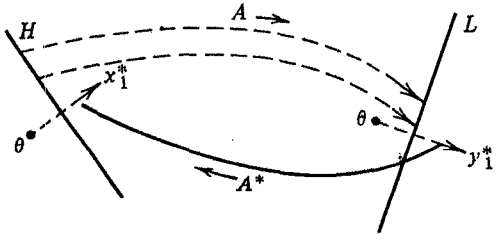

Figure 6.5 Geometric interpretation of adjoints

OPTIMIZATION IN HILBERT SPACE

Suppose A is a bounded linear operator from a Hilbert space G into a Hilbert space H; $A : G \to H$. Then, as pointed out previously, the linear equation $Ax = y$ may, for a given $y \in H$,

1. possess a unique solution $x \in G$,
2. possess no solution,
3. possess more than one solution.

Case 1 is in many respects the simplest. We found in Section 6.4 that in this case A has a unique bounded inverse A^{-1}. The other two cases are of interest in optimization since they allow some choice of an optimal x to be made. Indeed, most of the problems that were solved by the projection theorem can be viewed this way.

6.9 The Normal Equations

When no solution exists (case 2), we resolve the problem by finding an approximate solution.

Theorem 1. *Let G and H be Hilbert spaces and let $A \in B(G, H)$. Then for a fixed $y \in H$ the vector $x \in G$ minimizes $\|y - Ax\|$ if and only if $A^*Ax = A^*y$.*

Proof. The problem is obviously equivalent to that of minimizing $\|y - \hat{y}\|$ where $\hat{y} \in \mathscr{R}(A)$. Thus, by Theorem 1, Section 3.3 (the projection theorem without the existence part), $\hat{y}$ is a minimizing vector if and only if $y - \hat{y} \in [\mathscr{R}(A)]^{\perp}$. Hence, by Theorem 3 of Section 6.6, $y - \hat{y} \in \mathscr{N}(A^*)$. Or $\theta = A^*(y - \hat{y}) = A^*y - A^*Ax$. ∎

Theorem 1 is just a restatement of the first form of the projection theorem applied to the subspace $\mathscr{R}(A)$. There is no statement of existence in the theorem since in general $\mathscr{R}(A)$ may not be closed. Furthermore, there is no statement of uniqueness of the minimizing vector x since, although $\hat{y} = Ax$ is unique, the preimage of $\hat{y}$ may not be unique. If a unique solution always exists, i.e., if A^*A is invertible, the solution takes the form

$$x = (A^*A)^{-1}A^*y.$$

Example 1. We consider again the basic approximation problem in Hilbert space. Let $\{x_1, x_2, \ldots, x_n\}$ be an independent set of vectors in a real Hilbert space H. We seek the best approximation to $y \in H$ of the form $\hat{y} = \sum_{i=1}^{n} a_i x_i$.

Define the operator $A : E^n \to H$ by the equation

$$Aa = \sum_{i=1}^{n} a_i x_i,$$

where $a = (a_1, \ldots, a_n)$. The approximation problem is equivalent to minimizing $\|y - Aa\|$. Thus, according to Theorem 1, the optimal solution must satisfy $A^*Aa = A^*y$.

It remains to compute the operator A^*. Clearly, $A^*: H \to E^n$. For any $x \in H, a \in E^n$, we have

$$(x \mid Aa) = (x \mid \sum_{i=1}^{n} a_i x_i) = \sum_{i=1}^{n} a_i (x \mid x_i) = (z \mid a)_{E^n}$$

where $z = ((x \mid x_1), \ldots, (x \mid x_n))$. Thus, $A^*x = ((x \mid x_1), (x \mid x_2), \ldots, (x \mid x_n))$.

The operator A^*A maps E^n into E^n and is therefore represented by an $n \times n$ matrix. It is then easily deduced that the equation $A^*Aa = A^*y$ is equivalent to

$$\begin{bmatrix} (x_1 \mid x_1) & (x_2 \mid x_1) & \cdots & (x_n \mid x_1) \\ (x_1 \mid x_2) & & & \\ \vdots & & & \\ \vdots & & & \\ (x_1 \mid x_n) & & \cdots & (x_n \mid x_n) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} (y \mid x_1) \\ (y \mid x_2) \\ \vdots \\ \vdots \\ (y \mid x_n) \end{bmatrix},$$

the normal equations.

The familiar arguments for this problem show that the normal equations possess a unique solution and that the Gram matrix A^*A is invertible. Thus, $a = (A^*A)^{-1}A^*y$.

The above example illustrates that operator notation can streamline an optimization analysis by supplying a compact notational solution. The algebra required to compute adjoints and reduce the equations to expressions involving the original problem variables is, however, no shorter.

6.10 The Dual Problem

If the equation $Ax = y$ has more than one solution, we may choose the solution having minimum norm.

Theorem 1. *Let G and H be Hilbert spaces and let $A \in B(G, H)$ with range closed in H. Then the vector x of minimum norm satisfying $Ax = y$ is given by $x = A^*z$ where z is any solution of $AA^*z = y$.*

Proof. If x_1 is a solution of $Ax = y$, the general solution is $x = x_1 + u$ where $u \in \mathcal{N}(A)$. Since $\mathcal{N}(A)$ is closed, it follows that there exists a unique

vector x of minimum norm satisfying $Ax = y$ and that this vector is orthogonal to $\mathcal{N}(A)$. Thus, since $\mathcal{R}(A)$ is assumed closed,

$$x \in [\mathcal{N}(A)]^{\perp} = \mathcal{R}(A^*).$$

Hence $x = A^*z$ for some $z \in H$, and since $Ax = y$, we conclude that $AA^*z = y$. ∎

Note that if, as is frequently the case, the operator AA^* is invertible, the optimal solution takes the form

$$x = A^*(AA^*)^{-1}y.$$

Example 1. Suppose a linear dynamic system is governed by a set of differential equations of the form

$$\dot{x}(t) = Fx(t) + bu(t)$$

where x is an $n \times 1$ vector of time functions, F is an $n \times n$ matrix, b is an $n \times 1$ vector, and u is a scalar control function.

Assume that $x(0) = \theta$ and that it is desired to transfer the system to $x(T) = x_1$ by application of suitable control. Of the class of controls which accomplish the desired transfer, we seek the one of minimum energy $\int_0^T u^2(t)\,dt$. The problem includes the motor problem discussed in Chapter 3.

The explicit solution to the equation of motion is

$$x(T) = \int_0^T e^{F(T-t)}bu(t)\,dt.$$

Thus, defining the operator $A : L_2[0, T] \to E^n$ by

$$Au = \int_0^T e^{F(T-t)}bu(t)\,dt,$$

the problem is equivalent to that of determining the u of minimum norm satisfying $Au = x_1$.

Since $\mathcal{R}(A)$ is finite dimensional, it is closed. Thus the results of Theorem 1 apply and we write the optimal solution as

$$u = A^*z$$

where

$$AA^*z = x_1.$$

It remains to calculate the operators A^* and AA^*. For any $u \in L_2$, $y \in E^n$

$$(y \mid Au)_{E^n} = y' \int_0^T e^{F(T-t)}bu(t)\,dt = \int_0^T y'e^{F(T-t)}bu(t)\,dt$$

$$= (A^*y \mid u)_{L_2}$$

where

$$A^*y = b'e^{F'(T-t)}y.$$

Also, AA^* is the $n \times n$ matrix,

$$AA^* = \int_0^T e^{F(T-t)}bb'e^{F'(T-t)}\,dt.$$

If the matrix AA^* is invertible, the optimal control can be found as

$$u = A^*(AA^*)^{-1}x_1.$$

6.11 Pseudoinverse Operators

We now develop a more general and more complete approach to the problem of finding approximate or minimum norm solutions to $Ax = y$. The approach leads to the concept of the pseudoinverse of an operator A.

Suppose again that G and H are Hilbert spaces and that $A \in B(G, H)$ with $\mathscr{R}(A)$ closed. (In applications the closure of $\mathscr{R}(A)$ is usually supplied by the finite dimensionality of either G or H).

Definition. Among all vectors $x_1 \in G$ satisfying

$$\|Ax_1 - y\| = \min_x \|Ax - y\|,$$

let x_0 be the unique vector of minimum norm. The *pseudoinverse* $A^\dagger$ of A is the operator mapping y into its corresponding x_0 as y varies over H.

To justify the above definition, it must be verified that there is a unique x_0 corresponding to each $y \in H$. We observe first that $\min_x \|Ax - y\|$ is achieved since this amounts to approximating y by a vector in the closed subspace $\mathscr{R}(A)$. The approximation $\hat{y} = Ax_1$ is unique, although x_1 may not be.

The set of vectors x_1 satisfying $Ax_1 = \hat{y}$ is a linear variety, a translation of the subspace $\mathscr{N}(A)$. Thus, since this variety is closed, it contains a unique x_0 of minimum norm. Thus $A^\dagger$ is well defined. We show below that it is linear and bounded.

The above definition of the pseudoinverse is somewhat indirect and algebraic. We can develop a geometric interpretation of $A^\dagger$ so that certain of its properties become more transparent.

According to Theorem 1, Section 3.4, the space G can be expressed as

$$G = \mathscr{N}(A) \oplus \mathscr{N}(A)^\perp.$$

Likewise, since $\mathscr{R}(A)$ is assumed closed,

$$H = \mathscr{R}(A) \oplus \mathscr{R}(A)^\perp.$$

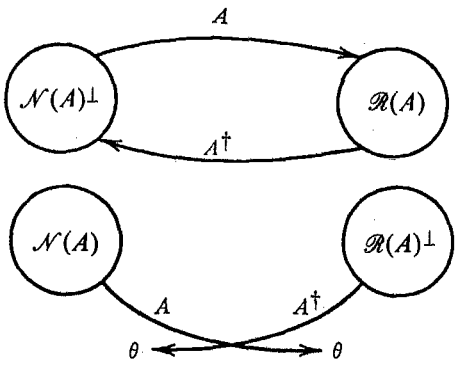

Figure 6.6 The pseudoinverse

The operator A restricted to $\mathcal{N}(A)^\perp$ can be regarded as an operator from the Hilbert space $\mathcal{N}(A)^\perp$ onto the Hilbert space $\mathcal{R}(A)$. Between these spaces A is one-to-one and onto and hence has a linear inverse which, according to the Banach inverse theorem, is bounded. This inverse operator defines $A^\dagger$ on $\mathcal{R}(A)$. Its domain is extended to all of H by defining $A^\dagger y = \theta$ for $y \in \mathcal{R}(A)^\perp$. Figure 6.6 shows this schematically and Figure 6.7 gives a geometric illustration of the relation of the various vectors in the problem.

It is easy to verify that this definition of $A^\dagger$ is in agreement with that of the last definition. Any $y \in H$ can be expressed uniquely as $y = \hat{y} + y_1$ where $\hat{y} \in \mathcal{R}(A)$, $y_1 \in \mathcal{R}(A)^\perp$. Thus $\hat{y}$ is the best approximation to y in $\mathcal{R}(A)$. Then $A^\dagger y = A^\dagger(\hat{y} + y_1) = A^\dagger \hat{y}$. Define $x_0 = A^\dagger y$. Then by definition $Ax_0 = \hat{y}$. Furthermore, $x_0 \in \mathcal{N}(A)^\perp$ and is therefore the minimum norm solution of $Ax_1 = \hat{y}$.

The pseudoinverse possesses a number of algebraic properties which are generalizations of corresponding properties for inverses. These properties are for the most part unimportant from our viewpoint of optimization; therefore they are not proved here but simply stated below.

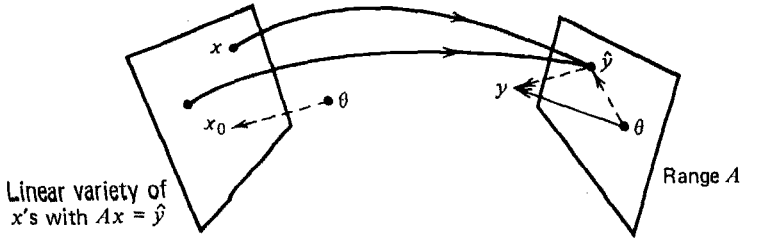

Figure 6.7 Relation between y and x_0

Proposition 1. *Let A be a bounded linear operator with closed range and let $A^\dagger$ denote its pseudoinverse. Then*

1. $A^\dagger$ *is linear.*
2. $A^\dagger$ *is bounded.*
3. $(A^\dagger)^\dagger = A$.
4. $(A^*)^\dagger = (A^\dagger)^*$.
5. $A^\dagger A A^\dagger = A^\dagger$.
6. $AA^\dagger A = A$.
7. $(A^\dagger A)^* = A^\dagger A$.
8. $A^\dagger = (A^*A)^\dagger A^*$.
9. $A^\dagger = A^*(AA^*)^\dagger$.

In certain limiting cases it is possible to give a simple explicit formula for $A^\dagger$. For instance, if A^*A is invertible, then $A^\dagger = (A^*A)^{-1}A^*$. If AA^* is invertible, then $A^\dagger = A^*(AA^*)^{-1}$. In general, however, a simple formula does not exist.

Example 1. The pseudoinverse arises in connection with approximation problems. Let $\{x_1, x_2, \ldots, x_n\}$ be a set of vectors in a Hilbert space H. In this example, however, these vectors are not assumed to be independent. As usual, we seek the best approximation of the form $\hat{y} = \sum_{i=1}^n a_i x_i$ to the vector y. Or by defining $A : E^n \to H$ by

$$Aa = \sum_{i=1}^{n} a_i x_i,$$

we seek the approximation to $Aa = y$. If the vector a achieving the best approximation is not unique, we then ask for the a of smallest norm which gives $\hat{y}$. Thus

$$a_0 = A^\dagger y.$$

The computation of $A^\dagger$ can be reduced by Proposition 1 to

$$A^\dagger = (A^*A)^\dagger A^*$$

so that the problem reduces to computing the pseudoinverse of the $n \times n$ Gram matrix $A^*A = G(x_1, x_2, \ldots, x_n)$.

6.12 Problems

1. Let $X = L_2[0, 1]$ and define A on X by

$$Ax = \int_0^1 K(t, s)x(s)\, ds$$

where

$$\int_0^1 \int_0^1 |K(t, s)|^2 \, dt \, ds < \infty.$$

Show that $A : X \to X$ and that A is bounded.

2. Let $X = L_p[0, 1]$, $1 < p < \infty$, $Y = L_q[0, 1]$, $1/p + 1/q = 1$. Define A by

$$Ax = \int_0^1 K(t, s)x(s) \, ds.$$

Show that $A \in B(X, Y)$ if $\int_0^1 \int_0^1 |K(t, s)|^q \, dt \, ds < \infty$.

3. Let A be a bounded linear operator from c_0 to l_∞. Show that corresponding to A there is an infinite matrix of scalars α_{ij}, $i, j = 1, 2, \ldots$, such that $y = Ax$ is expressed by the equations

$$\eta_i = \sum_{j=1}^{\infty} \alpha_{ij} \xi_j,$$

where $y = \{\eta_i\}$, $x = \{\xi_i\}$, and the norm of A is given by

$$\|A\| = \sup_i \sum_{j=1}^{\infty} |\alpha_{ij}|.$$

4. Prove the two-norm theorem: If X is a Banach space when normed by $\| \ \|_1$ and by $\| \ \|_2$ and if there is a constant c such that $\|x\|_1 \leq c \|x\|_2$ for all $x \in X$, then there is a constant C such that $\|x\|_2 \leq C \|x\|_1$ for all $x \in X$.
5. The *graph* of a transformation $T : X \to Y$ with domain $D \subset X$ is the set of points $(x, Tx) \in X \times Y$ with $x \in D$. Show that a bounded linear transformation with closed domain has a closed graph.
6. Let $X = C[a, b] = Y$ and let D be the subspace of X consisting of all continuously differentiable functions. Define the transformation T on D by $Tx = dx/dt$. Show that the graph of T is closed.
7. Prove the closed graph theorem: If X and Y are Banach spaces and T is a linear operator from X to Y with closed domain and closed graph, then T is bounded.
8. Show that a linear transformation mapping one Banach space into another is bounded if and only if its nullspace is closed.
9. Let H be the Hilbert space of n-tuples with inner product $(x \mid y)_Q = x'Qy$ where Q is a symmetric positive-definite matrix. Let an operator A on H be defined by an $n \times n$ matrix $[a_{ij}]$ in the usual sense. Find the matrix representation of A^*.

10. Let $X = L_p[0, 1]$, $1 < p < \infty$, $Y = L_q[0, 1]$, $1/p + 1/q = 1$. Let $A \in B(X, Y)$ be defined by $Ax = \int_0^1 K(t, s)x(s)\,ds$ where

$$\int_0^1 \int_0^1 |K(t, s)|^q \,dt\,ds < \infty.$$

(See Problem 2.) Find A^*.

11. Let X and Y be normed spaces and let $A \in B(X, Y)$. Show that $^\perp[\mathscr{R}(A^*)] = \mathscr{N}(A)$. (See Section 5.7.)
12. Let X, Y be Banach spaces and let $A \in B(X, Y)$ have closed range. Show that

$$\inf_{Ax=b} \|x\| = \max_{\|A^*y^*\| \le 1} \langle b, y^* \rangle.$$

Use this result to reinterpret the solution of the rocket problem of Example 3, Section 5.9.

13. Let X and Y be normed spaces and let G be the graph in $X \times Y$ of an operator $A \in B(X, Y)$. Show that $G^\perp$ is the graph of $-A^*$ in $X^* \times Y^*$.
14. Prove the Minkowski-Farkas lemma: Let A be an $m \times n$ matrix and b an n-dimensional vector. Then $Ax \le \theta$ implies $b'x \le 0$ if and only if $b = A'\lambda$ for some m-dimensional vector $\lambda \ge \theta$. Give a geometric interpretation of this result. (Inequalities among vectors are to be interpreted componentwise.)
15. Let M be a closed subspace of a Hilbert space H. The operator P defined by $Px = m$, where $x = m + n$ is the unique representation of $x \in H$ with $m \in M$, $n \in M^\perp$, is called the projection operator onto M. Show that a projection operator is linear and bounded with $\|P\| = 1$ if M is at least one dimensional.
16. Show that a bounded linear operator on a Hilbert space H is a projection operator if and only if:
    1. $P^2 = P$ (idempotent)
    2. $P^* = P$ (self-adjoint).
17. Two projection operators P_1 and P_2 on a Hilbert space are said to be orthogonal if $P_1P_2 = 0$. Show that two projection operators are orthogonal if and only if their ranges are orthogonal.
18. Show that the sum of two projection operators is a projection operator if and only if they are orthogonal.
19. Let G and H be Hilbert spaces and suppose $A \in B(G, H)$ with $\mathscr{R}(A)$ closed. Show that

$$A^\dagger = \lim_{\varepsilon \to 0+} (A^*A + \varepsilon I)^{-1}A^* = \lim_{\varepsilon \to 0+} A^*(AA^* + \varepsilon I)^{-1}$$

where the limits represent convergence in $B(H, G)$.

20. Find the pseudoinverse of the operator A on E^3 defined by the matrix

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

21. Let G, H, K be Hilbert spaces and let $B \in B(G, K)$ with range equal to K and $C \in B(K, H)$ with nullspace equal to $\{\theta\}$ (i.e., B is onto and C is one-to-one). Then for $A = CB$ we have $A^\dagger = B^\dagger C^\dagger$.

REFERENCES

§6.4. The Banach inverse theorem is intimately related to the closed graph theorem and the open mapping theorem. For further developments and applications, see Goffman and Pedrick [59].

§6.7. The concept of conjugate cone is related to various alternative concepts including dual cones, polar cones, and polar sets. See Edwards [46], Hurwicz [75], Fenchel [52], or Kelley, Namioka, et al. [85].

§6.10. See Balakrishnan [14].

§6.11. The concept of a pseudoinverse (or generalized inverse) was originally introduced for matrices by Moore [105], [106] and Penrose [117], [118]. See also Greville [67]. For an interesting introduction to the subject, see Albert [5]. Pseudoinverses on Hilbert space have been discussed by Tseng [146] and Ben Israel and Charnes [20]. Our treatment closely parallels that of Desoer and Whalen [38].

7

OPTIMIZATION OF FUNCTIONALS

7.1 Introduction

In previous chapters, while developing the elements of functional analysis, we often considered minimum norm optimization problems. Although the availability of a large variety of different norms provides enough flexibility for minimum norm problems to be of importance, there are many important optimization problems that cannot be formulated in these terms. In this chapter we consider optimization of more general objective functionals. However, much of the theory and geometric insight gained while studying minimum norm problems are of direct benefit in considering these more general problems.

Our study is guided by two basic geometric representations of nonlinear functionals. Each of these representations has its own particular advantages, and often it is enlightening to view a given situation in both ways. The first, and perhaps most obvious, geometric representation of a nonlinear functional is in terms of its graph. Suppose f is a functional defined on a subset D of the vector space X. The space X is then imbedded in the product space $R \times X$ where R is the real line. Elements of this space consist of ordered pairs (r, x). The graph of f is the surface in $R \times X$ consisting of the points $(f(x), x)$ with $x \in D$. Usually the R axis (i.e., points of the form (r, θ)) is regarded as the vertical axis, and the value of the functional at x is then regarded as the vertical distance of the graph above the point $x \in X$. In this representation a typical functional is visualized as in Figure 7.1. This representation is certainly familiar for one- or two-dimensional space. The point here is that it is convenient conceptually even in infinite-dimensional space.

The second representation is an extension of the technique of representing a linear functional by a hyperplane. A functional is described by its contours in the space X. A typical representation is shown in Figure 7.2. If f is sufficiently smooth, it is natural to construct hyperplanes tangent to the contours and to define the gradient of f. Again, this technique is

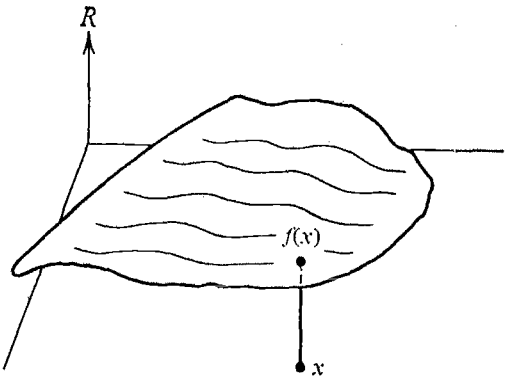

Figure 7.1 The graph of a functional

familiar in finite-dimensional space and is easily extended to infinite-dimensional space. The principal advantage of the second representation over the first is that for a given dimensionality of X, one less dimension is required for the representation.

The first half of this chapter deals with generalizations of the concepts of differentials, gradients, etc., to normed spaces and is essentially based on the second representation stated above. The detailed development is largely algebraic and manipulative in nature, although certain geometric interpretations are apparent. From this apparatus we obtain the local or variational theory of optimization paralleling the familiar theory in finite dimensions. The most elementary portion of the classical calculus of variations is used as a principal example of these methods.

The second half of the chapter deals with convex and concave functionals from which we obtain a global theory of optimization. This

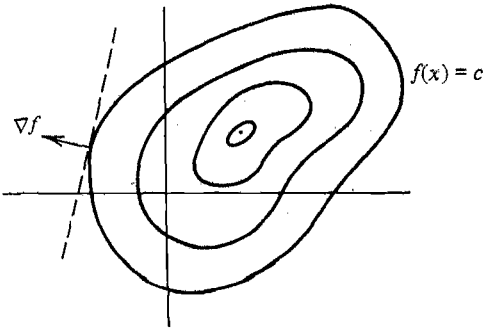

Figure 7.2 The contours of a functional

development, based essentially on the first representation for functionals, is largely geometric and builds on the theory of convex sets considered in Chapter 5. The interesting theory of conjugate functionals produces another duality theorem for a class of optimization problems.

LOCAL THEORY

7.2 Gateaux and Fréchet Differentials

In the following discussion let X be a vector space, Y a normed space, and T a (possibly nonlinear) transformation defined on a domain $D \subset X$ and having range $R \subset Y$.

Definition. Let $x \in D \subset X$ and let h be arbitrary in X. If the limit

$$\delta T(x; h) = \lim_{\alpha \to 0} \frac{1}{\alpha} [T(x + \alpha h) - T(x)] \tag{1}$$

exists, it is called the *Gateaux differential of T at x with increment h.* If the limit (1) exists for each $h \in X$, the transformation T is said to be *Gateaux differentiable at x.*

We note that it makes sense to consider the limit (1) only if $x + \alpha h \in D$ for all α sufficiently small. The limit (1) is, of course, taken in the usual sense of norm convergence in Y. We observe that for fixed $x \in D$ and h regarded as variable, the Gateaux differential defines a transformation from X to Y. In the particular case where T is linear, we have $\delta T(x; h) = T(h)$.

By far the most frequent application of this definition is in the case where Y is the real line and hence the transformation reduces to a (real-valued) functional on X. Thus if f is a functional on X, the Gateaux differential of f, if it exists, is

$$\delta f(x; h) = \frac{d}{d\alpha} f(x + \alpha h)\bigg|_{\alpha = 0},$$

and, for each fixed $x \in X$, $\delta f(x; h)$ is a functional with respect to the variable $h \in X$.

Example 1. Let $X = E^n$ and let $f(x) = f(x_1, x_2, \ldots, x_n)$ be a functional on E^n having continuous partial derivatives with respect to each variable x_i. Then the Gateaux differential of f is

$$\delta f(x; h) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} h_i.$$

Example 2. Let $X = C[0, 1]$ and let $f(x) = \int_0^1 g(x(t), t)\, dt$, where it is assumed that the function g_x exists and is continuous with respect to x and t. Then

$$\delta f(x; h) = \frac{d}{d\alpha} \int_0^1 g(x(t) + \alpha h(t), t)\, dt \Big|_{\alpha=0}$$

Interchange of the order of differentiation and integration is permissible under our assumptions on g and hence

$$\delta f(x; h) = \int_0^1 g_x(x, t) h(t)\, dt.$$

Example 3. If $X = E^n$, $Y = E^m$, and T is a continuously differentiable mapping of X into Y, then $\delta T(x; h)$ exists. It is the vector in Y equal to the vector $h \in X$ multiplied by the matrix made up of partial derivatives of T at x.

The Gateaux differential generalizes the concept of directional derivative familiar in finite-dimensional space. The existence of the Gateaux differential is a rather weak requirement, however, since its definition requires no norm on X; hence, properties of the Gateaux differential are not easily related to continuity. When X is normed, a more satisfactory definition is given by the Fréchet differential.

Definition. Let T be a transformation defined on an open domain D in a normed space X and having range in a normed space Y. If for fixed $x \in D$ and each $h \in X$ there exists $\delta T(x; h) \in Y$ which is linear and continuous with respect to h such that

$$\lim_{\|h\| \to 0} \frac{\|T(x + h) - T(x) - \delta T(x; h)\|}{\|h\|} = 0,$$

then T is said to be *Fréchet differentiable at* x and $\delta T(x; h)$ is said to be the *Fréchet differential of* T *at* x *with increment* h.

We use the same symbol for the Fréchet and Gateaux differentials since generally it is apparent from the context which is meant.

Proposition 1. *If the transformation* T *has a Fréchet differential, it is unique.*

Proof. Suppose both $\delta T(x; h)$ and $\delta' T(x; h)$ satisfy the requirements of the last definition. Then

$$\|\delta T(x; h) - \delta' T(x; h)\| \leq \|T(x + h) - T(x) - \delta T(x; h)\| + \|T(x + h) - T(x) - \delta' T(x; h)\|$$

or $\|\delta T(x; h) - \delta' T(x; h)\| = o(\|h\|)$. Since $\delta T(x; h) - \delta' T(x; h)$ is bounded and linear in h, it must be zero. ∎

Proposition 2. *If the Fréchet differential of T exists at x, then the Gateaux differential exists at x and they are equal.*

Proof. Denote the Fréchet differential by $\delta T(x; h)$. By definition we have for any h,

$$\frac{1}{\alpha}\|T(x+\alpha h) - T(x) - \delta T(x; \alpha h)\| \to 0, \qquad \text{as } \alpha \to 0.$$

Thus, by the linearity of $\delta T(x; \alpha h)$ with respect to α,

$$\lim_{\alpha \to 0} \frac{T(x+\alpha h) - T(x)}{\alpha} = \delta T(x; h). \quad \blacksquare$$

A final property is given by the following proposition.

Proposition 3. *If the transformation T defined on an open set D in X has a Fréchet differential at x, then T is continuous at x.*

Proof. Given $\varepsilon > 0$, there is a sphere about x such that for $x + h$ in this sphere

$$\|T(x+h) - T(x) - \delta T(x; h)\| < \varepsilon \|h\|.$$

Thus $\|T(x+h) - T(x)\| < \varepsilon\|h\| + \|\delta T(x; h)\| < M\|h\|$. $\blacksquare$

We are primarily concerned with Fréchet differentials rather than Gateaux differentials and often assume their existence or, equivalently, assume the satisfaction of conditions which are sufficient to imply their existence.

Example 4. We show that the differential

$$\delta f(x; h) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} h_i$$

in Example 1 (of the functional f on E^n) is a Fréchet differential. Obviously, $\delta f(x; h)$ is linear and continuous in h. It is therefore only necessary to verify the basic limit property required of Fréchet differentials.

Given $\varepsilon > 0$, the continuity of the partial derivatives implies that there is a neighborhood $S(x; \delta)$ such that

$$\left| \frac{\partial f(x)}{\partial x_i} - \frac{\partial f(y)}{\partial x_i} \right| < \frac{\varepsilon}{n}$$

for all $y \in S(x; \delta)$ and $i = 1, 2, \ldots, n$.

Define the unit vectors e_i in the usual way, $e_i = (0, 0, \ldots, 1, 0, \ldots, 0)$, and for $h = \sum_{i=1}^{n} h_i e_i$ define $g_0 = \theta$ and

$$g_k = \sum_{i=1}^{k} h_i e_i, \qquad \text{for } k = 1, 2, \ldots, n.$$

We note that $\|g_k\| \leq \|h\|$ for all k. Then

$$\left|f(x+h) - f(x) - \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} h_i\right| = \left|\sum_{k=1}^{n} \left\{f(x+g_k) - f(x+g_{k-1}) - \frac{\partial f}{\partial x_k} h_k\right\}\right|$$

$$\leq \sum_{k=1}^{n} \left|f(x+g_k) - f(x+g_{k-1}) - \frac{\partial f}{\partial x_k} h_k\right|.$$

Now we examine the k-th term in the above summation. The vector $x + g_k$ differs from $x + g_{k-1}$ only in the k-th component. In fact, $x + g_k = x + g_{k-1} + h_k e_k$. Thus, by the mean value theorem for functions of a single variable,

$$f(x+g_k) - f(x+g_{k-1}) = \frac{\partial f}{\partial x_k}(x + g_{k-1} + \alpha e_k)h_k$$

for some α, $0 \leq \alpha \leq h_k$.

Also, $x + g_{k-1} + \alpha e_k \in S(x; \delta)$ if $\|h\| < \delta$. Thus

$$\left|f(x+g_k) - f(x+g_{k-1}) - \frac{\partial f(x)}{\partial x_k} h_k\right| < \frac{\varepsilon}{n}\|h\|.$$

Finally, it follows that

$$\left|f(x+h) - f(x) - \sum_{i=1}^{n} \frac{\partial f}{\partial x_k} h_i\right| < \varepsilon\|h\|$$

for all h, $\|h\| < \delta$.

Example 5. We show that the differential

$$\delta f(x; h) = \int_0^1 g_x(x, t)h(t)\,dt$$

in Example 2 is a Fréchet differential. We have

$$|f(x+h) - f(x) - \delta f(x; h)| = \left|\int_0^1 \{g(x+h, t) - g(x, t) - g_x(x, t)h(t)\}\,dt\right|.$$

For a fixed t we have, by the one-dimensional mean value theorem,

$$g(x(t) + h(t), t) - g(x(t), t) = g_x(\bar{x}(t), t)h(t)$$

where $|x(t) - \bar{x}(t)| \leq |h(t)|$. Given $\varepsilon > 0$, the uniform continuity of g_x in x and t implies that there is a $\delta > 0$ such that for $\|h\| < \delta$, $|g_x(x+h, t) - g_x(x, t)| < \varepsilon$. Therefore, we have

$$|f(x+h) - f(x) - \delta f(x, h)| = \left|\int_0^1 (g_x(\bar{x}, t) - g_x(x, t))h(t)\,dt\right| \leq \varepsilon\|h\|$$

for $\|h\| < \delta$. The result follows.

Example 6. Let $X = C^n[0, 1]$, the space of continuous n-vector functions on $[0, 1]$. Let $Y = C^m[0, 1]$ and define $T: X \to Y$ by

$$T(x) = \int_0^t F[x(\tau), \tau]\, d\tau$$

where $F = (f_1, f_2, \ldots, f_m)$ has continuous partial derivatives with respect to its arguments. The Gateaux differential of T is easily seen to be

$$\delta T(x; h) = \int_0^t F_x[x(\tau), \tau]h(\tau)\, d\tau.$$

By combining the analyses of Examples 4 and 5, we can conclude that this is actually a Fréchet differential. Note also that since the partial derivatives of F are continuous, $\delta T(x; h)$ is continuous in the variable x.

7.3 Fréchet Derivatives

Suppose that the transformation T defined on an open domain $D \subset X$ is Fréchet differentiable throughout D. At a fixed point $x \in D$ the Fréchet differential $\delta T(x; h)$ is then, by definition, of the form $\delta T(x; h) = A_x h$, where A_x is a bounded linear operator from X to Y. Thus, as x varies over D, the correspondence $x \to A_x$ defines a transformation from D into the normed linear space $B(X, Y)$; this transformation is called the *Fréchet derivative* T' *of* T. Thus we have, by definition, $\delta T(x; h) = T'(x)h$.

If the correspondence $x \to T'(x)$ is continuous at the point x_0 (i.e., if given $\varepsilon > 0$ there is $\delta > 0$ such that $\|x - x_0\| < \delta$ implies $\|T'(x) - T'(x_0)\| < \varepsilon$), we say that the Fréchet derivative of T is *continuous* at x_0. This should not be confused with the statement that $T'(x_0)$ is a continuous mapping from X to Y, a property that is basic to the definition of the Fréchet derivative. If the derivative of T is continuous on some open sphere S, we say that T is *continuously Fréchet differentiable* on S. All of the examples of Fréchet differentiable transformations in the last section are in fact continuously Fréchet differentiable.

In the special case where the original transformation is simply a functional f on the space X, we have $\delta f(x; h) = f'(x)h$ where $f'(x) \in X^*$ for each x. The element $f'(x)$ is called the gradient of f at x and is sometimes denoted $\nabla f(x)$ rather than $f'(x)$. We sometimes write $\langle h, f'(x)\rangle$ for $\delta f(x; h)$ since $f'(x) \in X^*$, but usually we prefer $f'(x)h$ which is consistent with the notation for differentials of arbitrary transformations.

Much of the theory of ordinary derivatives can be generalized to Fréchet derivatives. For instance, the implicit function theorem and Taylor series have very satisfactory extensions. The interested reader should consult the

references cited at the end of the chapter. In this section we discuss the elementary properties of Fréchet derivatives used in later sections.

It follows immediately from the definition that if T_1 and T_2 are Fréchet differentiable at $x \in D$, then $\alpha_1 T_1 + \alpha_2 T_2$ is Fréchet differentiable at x and $(\alpha_1 T_1 + \alpha_2 T_2)'(x) = \alpha_1 T_1'(x) + \alpha_2 T_2'(x)$. We next show that the chain rule applies to Fréchet derivatives.

Proposition 1. *Let S be a transformation mapping an open set $D \subset X$ into an open set $E \subset Y$ and let P be a transformation mapping E into a normed space Z. Put $T = PS$ and suppose S is Fréchet differentiable at $x \in D$ and P is Fréchet differentiable at $y = S(x) \in E$. Then T is Fréchet differentiable at x and $T'(x) = P'(y)S'(x)$.*

Proof. For $h \in X$, $x + h \in D$, we have

$$T(x + h) - T(x) = P[S(x + h)] - P[S(x)] = P(y + g) - P(y)$$

where $g = S(x + h) - S(x)$. Thus $\|T(x + h) - T(x) - P'(y)g\| = o(\|g\|)$. Since, however,

$$\|g - S'(x)h\| = o(\|h\|),$$

we obtain

$$\|T(x + h) - T(x) - P'(y)S'(x)h\| = o(\|h\|) + o(\|g\|).$$

Since, according to Proposition 3 of Section 7.2, S is continuous at x, we conclude that $\|g\| = O(\|h\|)$ and hence

$$T'(x)h = P'(y)S'(x)h. \quad \blacksquare$$

We now give a very useful inequality that replaces the mean value theorem for ordinary functions.

Proposition 2. *Let T be Fréchet differentiable on an open domain D. Let $x \in D$ and suppose that $x + \alpha h \in D$ for all α, $0 \leq \alpha \leq 1$. Then*

$$\|T(x + h) - T(x)\| \leq \|h\| \sup_{0 < \alpha < 1} \|T'(x + \alpha h)\|.$$

Proof. Let y^* be a nonzero element of Y^* aligned with the element $T(x + h) - T(x)$. The function $\varphi(\alpha) = y^*[T(x + \alpha h)]$ is defined on the interval $[0, 1]$ and, by the chain rule, has derivative

$$\varphi'(\alpha) = y^*[T'(x + \alpha h)h].$$

By the mean value theorem for functions of a real variable, we have

$$\varphi(1) - \varphi(0) = \varphi'(\alpha_0), \qquad 0 < \alpha_0 < 1,$$

and hence

$$|y^*[T(x+h)-T(x)]| \leq \|y^*\| \sup_{0<\alpha<1} \|T'(x+\alpha h)\|\|h\|,$$

and since y^* is aligned with $T(x+h)-T(x)$,

$$\|T(x+h)-T(x)\| \leq \|h\| \sup_{0<\alpha<1} \|T'(x+\alpha h)\|. \blacksquare$$

If $T: X \to Y$ is Fréchet differentiable on an open domain $D \subset X$, the derivative T' maps D into $B(X, Y)$ and may itself be Fréchet differentiable on a subset $D_1 \subset D$. In this case the Fréchet derivative of T' is called the second Fréchet derivative of T and is denoted by T''.

Example 1. Let f be a functional on $X = E^n$ having continuous partial derivatives up to second order. Then $f''(x_0)$ is an operator from E^n to E^n having matrix form

$$f''(x_0) = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right]_{x=x_0},$$

where x_i is the i-th component of x.

The following inequality can be proved in a manner paralleling that of Proposition 2.

Proposition 3. *Let T be twice Fréchet differentiable on an open domain D. Let $x \in D$ and suppose that $x + \alpha h \in D$ for all α, $0 \leq \alpha \leq 1$. Then*

$$\|T(x+h) - T(x) - T'(x)h\| \leq \tfrac{1}{2}\|h\|^2 \sup_{0<\alpha<1} \|T''(x+\alpha h)\|.$$

7.4 Extrema

It is relatively simple to apply the concepts of Gateaux and Fréchet differentials to the problem of minimizing or maximizing a functional on a linear space. The technique leads quite naturally to the rudiments of the calculus of variations where, in fact, the abstract concept of differentials originated. In this section, we extend the familiar technique of minimizing a function of a single variable by ordinary calculus to a similar technique based on more general differentials. In this way we obtain analogs of the classical necessary conditions for local extrema and, in a later section, the Lagrange technique for constrained extrema.

Definition. Let f be a real-valued functional defined on a subset Ω of a normed space X. A point $x_0 \in \Omega$ is said to be a *relative minimum* of f on Ω if there is an open sphere N containing x_0 such that $f(x_0) \leq f(x)$ for all $x \in \Omega \cap N$. The point x_0 is said to be a *strict relative minimum* of f on Ω

if $f(x_0) < f(x)$ for all $x \neq x_0$, $x \in \Omega \cap N$. *Relative maxima* are defined similarly.

We use the term *extremum* to refer to either a maximum or a minimum over any set. A *relative extremum* (over a subset of a normed space) is also referred to as a *local extremum*. The set Ω on which an extremum problem is defined is sometimes called the *admissible set*.

Theorem 1. *Let the real-valued functional f have a Gateaux differential on a vector space X. A necessary condition for f to have an extremum at $x_0 \in X$ is that $\delta f(x_0; h) = 0$ for all $h \in X$.*

Proof. For every $h \in X$, the function $f(x_0 + \alpha h)$ of the real variable α must achieve an extremum at $\alpha = 0$. Thus, by the ordinary calculus,

$$\frac{d}{d\alpha} f(x_0 + \alpha h)\bigg|_{\alpha=0} = 0. \blacksquare$$

A point at which $\delta f(x; h) = 0$ for all h is called a *stationary point*; hence, the above theorem merely states that extrema occur at stationary points. It should be noted that a similar result holds for a local extremum of a functional defined on an open subset of a normed space, the proof being identical for both cases.

The simplicity of Theorem 1 can be misleading for it is a result of great utility. Much of the calculus of variations can be regarded as a simple consequence of this one result. Indeed, many interesting problems are solved by careful identification of an appropriate vector space X and some algebraic manipulations to obtain the differential. There are a number of useful generalizations of Theorem 1. We offer one of the simplest.

Theorem 2. *Let f be a real-valued functional defined on a vector space X. Suppose that x_0 minimizes f on the convex set $\Omega \subset X$ and that f is Gateaux differentiable at x_0. Then*

$$\delta f(x_0; x - x_0) \geq 0$$

for all $x \in \Omega$.

Proof. Since Ω is convex, $x_0 + \alpha(x - x_0) \in \Omega$ for $0 \leq \alpha \leq 1$ and hence, by ordinary calculus,

$$\frac{d}{d\alpha} f(x_0 + \alpha(x - x_0))\bigg|_{\alpha=0} \geq 0$$

for a minimum at x_0. $\blacksquare$

*7.5 Euler-Lagrange Equations

The classical problem in the calculus of variations is that of finding a function x on the interval $[t_1, t_2]$ minimizing an integral functional of the form

$$J = \int_{t_1}^{t_2} f[x(t), \dot{x}(t), t]\, dt.$$

To specify the problem completely, we must agree on the class of functions within which we seek the extremum—the so-called admissible set. We assume that the function f is continuous in x, $\dot{x}$, and t and has continuous partial derivatives with respect to x and $\dot{x}$. We seek a solution in the space $D[t_1, t_2]$. In the simplest version of the problem, we assume that the end points $x(t_1)$ and $x(t_2)$ are fixed. This further restricts the admissible set.

Starting with a given admissible vector x, we consider vectors of the form $x + h$ that are admissible. The class of such vectors h is called the class of *admissible variations*. In the case of fixed end points, it is clear that the class of admissible variations is the subspace of $D[t_1, t_2]$, consisting of functions which vanish at t_1 and t_2. The necessary condition for the extremum problem is that for all such h, $\delta J(x; h) = 0$.

The differential of J is

$$\delta J(x; h) = \frac{d}{d\alpha} \int_{t_1}^{t_2} f(x + \alpha h, \dot{x} + \alpha \dot{h}, t)\, dt \bigg|_{\alpha = 0}$$

or, equivalently,

$$\delta J(x; h) = \int_{t_1}^{t_2} f_x(x, \dot{x}, t) h(t)\, dt + \int_{t_1}^{t_2} f_{\dot{x}}(x, \dot{x}, t) \dot{h}(t)\, dt, \tag{1}$$

and it is easily verified that this differential is actually Fréchet. Equating this differential to zero and assuming that the function $f_{\dot{x}}$ has a continuous derivative with respect to t when the optimal solution is substituted for x, we may integrate by parts to obtain

$$\delta J(x; h) = \int_{t_1}^{t_2} [f_x(x, \dot{x}, t) - \frac{d}{dt} f_{\dot{x}}(x, \dot{x}, t)] h(t)\, dt + f_{\dot{x}}(x, \dot{x}, t) h(t) \bigg|_{t_1}^{t_2} = 0.$$

The boundary terms vanish for admissible h and thus the necessary condition is

$$\int_{t_1}^{t_2} \left[f_x(x, \dot{x}, t) - \frac{d}{dt} f_{\dot{x}}(x, \dot{x}, t) \right] h(t)\, dt = 0$$

for all $h \in D[t_1, t_2]$ vanishing at t_1 and t_2.

Since the term multiplying $h(t)$ in the integrand is continuous, it readily follows (see Lemma 1 below) that it must vanish identically on $[t_1, t_2]$. Thus we conclude that the extremal x must satisfy the Euler-Lagrange equation

$$f_x(x, \dot{x}, t) - \frac{d}{dt} f_{\dot{x}}(x, \dot{x}, t) = 0. \tag{2}$$

The above derivation of the Euler-Lagrange equations suffers from the weakness of the assumption that

$$\frac{d}{dt} f_{\dot{x}}$$

is continuous at the optimal solution. Actually we have no basis on which to assume that the solution is in fact smooth enough for this assumption to hold. The alternate derivation, given after the following three lemmas, avoids this drawback.

Lemma 1. *If $\alpha(t)$ is continuous on $[t_1, t_2]$ and $\int_{t_1}^{t_2} \alpha(t)h(t)\,dt = 0$ for every $h \in D[t_1, t_2]$ with $h(t_1) = h(t_2) = 0$, then $\alpha(t) \equiv 0$ on $[t_1, t_2]$.*

Proof. Assume that $\alpha(t)$ is nonzero, say positive, for some $t \in [t_1, t_2]$. Then $\alpha(t)$ is positive on some interval $[t_1', t_2'] \subset [t_1, t_2]$. Let

$$h(t) = \begin{cases} (t - t_1')^2(t_2' - t)^2 & t \in [t_1', t_2'] \\ 0 & \text{otherwise.} \end{cases}$$

The function h satisfies the hypotheses of the lemma and

$$\int_{t_1}^{t_2} \alpha(t)h(t)\,dt > 0. \quad \blacksquare$$

Lemma 2. *If $\alpha(t)$ is continuous in $[t_1, t_2]$ and $\int_{t_1}^{t_2} \alpha(t)\dot{h}(t)\,dt = 0$ for every $h \in D[t_1, t_2]$ with $h(t_1) = h(t_2) = 0$, then $\alpha(t) \equiv c$ in $[t_1, t_2]$ where c is a constant.*

Proof. Let c be the unique constant satisfying $\int_{t_1}^{t_2} [\alpha(t) - c]\,dt = 0$ and let

$$h(t) = \int_{t_1}^{t} [\alpha(\tau) - c]\,d\tau.$$

Then

$$\begin{aligned} \int_{t_1}^{t_2} [\alpha(t) - c]^2\,dt &= \int_{t_1}^{t_2} [\alpha(t) - c)]\dot{h}(t)\,dt \\ &= \int_{t_1}^{t_2} \alpha(t)\dot{h}(t)\,dt - c[h(t_2) - h(t_1)] = 0, \end{aligned}$$

and hence $\alpha(t) \equiv c$. $\blacksquare$

Lemma 3. *If $\alpha(t)$ and $\beta(t)$ are continuous in $[t_1, t_2]$ and*

$$\int_{t_1}^{t_2} [\alpha(t)h(t) + \beta(t)\dot{h}(t)]\, dt = 0 \tag{3}$$

for every $h \in D[t_1, t_2]$ with $h(t_1) = h(t_2) = 0$, then β is differentiable and $\dot{\beta}(t) \equiv \alpha(t)$ in $[t_1, t_2]$.

Proof. Define

$$A(t) = \int_{t_1}^{t} \alpha(\tau)\, d\tau.$$

Then by integration by parts we have

$$\int_{t_1}^{t_2} \alpha(t)h(t)\, dt = -\int_{t_1}^{t_2} A(t)\dot{h}(t)\, dt.$$

Therefore (3) becomes

$$\int_{t_1}^{t_2} [-A(t) + \beta(t)]\dot{h}(t)\, dt = 0$$

which by Lemma 2 implies

$$\beta(t) = A(t) + c$$

for some constant c. Hence, by the definition of A, $\dot{\beta}(t) = \alpha(t)$. ∎

Now, in view of Lemma 3, it is clear that the Euler-Lagrange equation (2) follows directly from equation (1) without an *a priori* assumption of the differentiability of $f_{\dot{x}}$.

Example 1. (Minimum Arc Length) Given t_1, t_2 and $x(t_1)$, $x(t_2)$, let us employ the Euler-Lagrange equations to determine the curve in $D[t_1, t_2]$ connecting these points with minimum arc length. We thus seek to minimize

$$J = \int_{t_1}^{t_2} \sqrt{1 + (\dot{x})^2}\, dt.$$

From (2) we obtain immediately

$$\frac{d}{dt}\frac{\partial}{\partial \dot{x}} \sqrt{1 + (\dot{x})^2} = 0$$

or, equivalently,

$$\dot{x} = \text{const.}$$

Thus the extremizing arc is the straight line connecting the two points.

Example 2. (Estate Planning) What is the lifetime plan of investment and expenditure that maximizes total enjoyment for a man having a fixed quantity of savings S? We assume that the man has no income other than that obtained through his investment. His rate of enjoyment (or utility) at a given time is a certain function U of r, his rate of expenditure. Thus, we assume that it is desired to maximize

$$\int_0^T e^{-\beta t} U[r(t)]\, dt$$

where the $e^{-\beta t}$ term reflects the notion that future enjoyment is counted less today.

If $x(t)$ is the total capital at time t, then

$$\dot{x}(t) = \alpha x(t) - r(t)$$

where α is the interest rate of investment. Thus the problem is to maximize

$$\int_0^T e^{-\beta t} U[\alpha x(t) - \dot{x}(t)]\, dt$$

subject to $x(0) = S$ and $x(T) = 0$ (or some other fixed value). Of course, there is the additional constraint $x(t) \geq 0$, but, in the cases we consider, this turns out to be satisfied automatically.

From the Euler-Lagrange equation (2), we obtain

$$\alpha e^{-\beta t} U'[\alpha x(t) - \dot{x}(t)] + \frac{d}{dt} e^{-\beta t} U'[\alpha x(t) - \dot{x}(t)] = 0$$

where U' is the derivative of the utility function U. This becomes

$$\frac{d}{dt} U'[\alpha x(t) - \dot{x}(t)] = (\beta - \alpha) U'[\alpha x(t) - \dot{x}(t)]. \tag{4}$$

Hence, integrating (4), we find that

$$U'[r(t)] = U'[r(0)] e^{(\beta - \alpha)t}. \tag{5}$$

Hence, the form of the time dependence of r can be obtained explicitly once U is specified.

A simple utility function, which reflects both the intuitive notions of diminishing marginal enjoyment (i.e., U' decreasing) and infinite marginal enjoyment at zero expenditure (i.e., $U'[0] = \infty$), is $U[r] = 2r^{1/2}$. Substituting this in (5), we obtain

$$\alpha x(t) - \dot{x}(t) = r(t) = r(0) e^{2(\alpha - \beta)t}.$$

Integrating this last equation, we have

$$x(t) = e^{\alpha t}x(0) + \frac{r(0)}{\alpha - 2\beta}\left[e^{\alpha t} - e^{2(\alpha-\beta)t}\right]$$

(6)

$$= \left\{x(0) - \frac{r(0)}{2\beta - \alpha}\right\}e^{\alpha t} + \frac{r(0)}{2\beta - \alpha}\, e^{2(\alpha-\beta)t}.$$

Assuming $\alpha > \beta > \alpha/2$, we may find $r(0)$ from (6) by requiring $x(T) = 0$. Thus

$$r(0) = \frac{(2\beta - \alpha)x(0)}{1 - e^{-(2\beta-\alpha)T}}.$$

The total capital grows initially and then decreases to zero.

*7.6 Problems with Variable End Points

The class of admissible functions for a given problem is not always a linear variety of functions nor even a convex set. Such problems may, however, generally be approached in essentially the same manner as before. Basically the technique is to define a continuously differentiable one-parameter family of admissible curves that includes the optimal curve as a member. A necessary condition is obtained by equating the derivative of the objective functional, with respect to the parameter, equal to zero. In the case where the admissible class is a linear variety, we consider the family $x + \varepsilon h$ and differentiate with respect to ε. In other cases, a more general family $x(\varepsilon)$ of admissible curves is considered.

An interesting class of problems that can be handled in this way is calculus of variations problems having variable end points. Specifically, suppose we seek an extremum of the functional

$$J = \int_{t_1}^{t_2} f(x, \dot{x}, t)\, dt$$

where the interval $[t_1, t_2]$, as well as the function $x(t)$, must be chosen. In a typical problem the two end points are constrained to lie on fixed curves in the $x - t$ plane. We assume here that the left end point is fixed and allow the right end point to vary along a curve S described by the function $x = g(t)$, as illustrated in Figure 7.3.

Suppose $x(\varepsilon, t)$ is a one-parameter family of functions emanating from the point 1 and terminating on the curve S. The termination point is $x(\varepsilon, t_2(\varepsilon))$ and satisfies $x(\varepsilon, t_2(\varepsilon)) = g(t_2(\varepsilon))$. We assume that the family is defined for $\varepsilon \in [-a, a]$ for some $a > 0$ and that the desired extremal is the

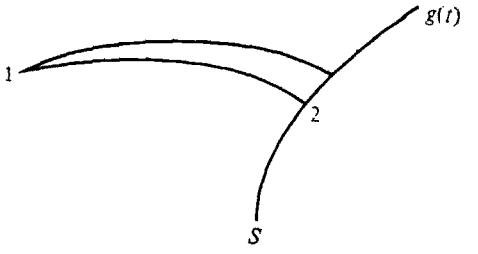

Figure 7.3 A variable end-point problem

curve corresponding to $\varepsilon = 0$. The variation δJ of the functional J is defined as the first derivative of J with respect to ε and must vanish at $\varepsilon = 0$.

Defining, in the obvious way,

$$\delta x(t) = \frac{d}{d\varepsilon} x(\varepsilon, t)\bigg|_{\varepsilon=0}$$

$$\delta t_2 = \frac{d}{d\varepsilon} t_2(\varepsilon)\bigg|_{\varepsilon=0},$$

we have

$$J(\varepsilon) = \int_{t_1}^{t_2(\varepsilon)} f[x(t, \varepsilon), \dot{x}(t, \varepsilon), t]\, dt$$

$$\delta J = f[x(t_2), \dot{x}(t_2), t_2]\delta t_2 + \int_{t_1}^{t_2} \{f_x(x, \dot{x}, t)\delta x + f_{\dot{x}}(x, \dot{x}, t)\delta\dot{x}\}\, dt = 0.$$

Arguments similar to those of the last section lead directly to the necessary conditions

$$f_x(x, \dot{x}, t) = \frac{d}{dt} f_{\dot{x}}(x, \dot{x}, t) \tag{1}$$

and

$$f[x(t_2), \dot{x}(t_2), t_2]\delta t_2 + f_{\dot{x}}[x(t_2), \dot{x}(t_2), t_2]\delta x(t_2) = 0. \tag{2}$$

Condition (1) is again the Euler-Lagrange equation, but in addition we have the transversality condition (2) which must be employed to determine the termination point. The transversality condition (2) must hold for all admissible $\delta x(t_2)$ and δt_2. These two quantities are not independent, however, since

$$x[\varepsilon, t_2(\varepsilon)] = g[t_2(\varepsilon)]. \tag{3}$$

Upon differentiating (3) with respect to ε, we obtain the relation

$$\delta x(t_2) + \dot{x}(t_2)\delta t_2 = \dot{g}(t_2)\delta t_2$$

and hence the complete set of necessary conditions is (1) and the transversality condition

$$\{f(x, \dot{x}, t) + [\dot{g} - \dot{x}]f_{\dot{x}}(x, \dot{x}, t)\}\big|_{t=t_2} = 0. \tag{4}$$

Example 1. As an example of the transversality condition, we consider the simple problem of finding the differentiable curve of minimum arc length from the origin (0, 0) to the curve g. Thus we seek to extremize the functional

$$J = \int_0^{t_2} \sqrt{1 + (\dot{x})^2}\, dt,$$

where t_2 is the point of intersection of the line.

As in Example 1 of the last section, the Euler-Lagrange equation leads to $\dot{x} = \text{const}$ and, hence, the extremal must be a straight line. The transversality condition in this case is

$$\left\{\sqrt{1 + (\dot{x})^2} + (\dot{g} - \dot{x})\frac{\dot{x}}{\sqrt{1 + (\dot{x})^2}}\right\}\Bigg|_{t=t_2} = 0,$$

or

$$\dot{x}(t_2) = -\frac{1}{\dot{g}(t_2)}.$$

Thus the extremal arc must be orthogonal to the tangent of g at t_2. (See Figure 7.4.)

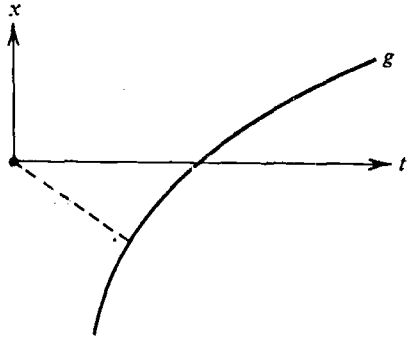

Figure 7.4 Minimum distance to a curve

7.7 Problems with Constraints

In most optimization problems the optimal vector is required to satisfy some type of constraint. In the simplest calculus of variations problem, for instance, the end points of the curve are constrained to fixed points.

Such problems, and those of greater complexity, often can be resolved (in the sense of establishing necessary conditions) by considering a one-parameter family of vectors satisfying the constraints which contains the optimal vector as a member. A complication arises, however, when the constraint set is defined implicitly in terms of a set of functional equations rather than explicitly as a constraint surface in the space. In such cases the one-parameter family must be constructed implicitly. In this section we carry out the necessary construction for a finite number of functional constraints and arrive at our first Lagrange multiplier theorem.

A more general discussion of constrained optimization problems, including the material of this section as a special case, is given in Chapter 9. The later discussion parallels, but is much deeper than, the one given here.

The problem investigated in this section is that of optimizing a functional f subject to n nonlinear constraints given in the implicit form:

$$\begin{aligned} g_1(x) &= 0 \\ g_2(x) &= 0 \\ &\vdots \\ g_n(x) &= 0. \end{aligned} \tag{1}$$

These n equations define a region Ω in the space X within which the optimal vector x_0 is constrained to lie. Throughout this section it is assumed that all functionals f, g_i are continuous and Fréchet differentiable on the normed space X.

Before embarking on the general resolution of the problem, let us briefly consider the geometry of the problem. The case where X is two dimensional is depicted in Figure 7.5. If x_0 is optimal, the functional f has an extremum at x_0 with respect to small displacements along Ω. Under sufficient smoothness conditions, it seems that f has an extremum at x_0 with respect to small displacements along T, the tangent to Ω at x_0.

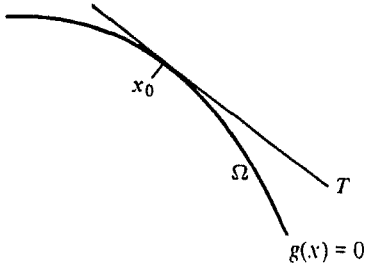

Figure 7.5 Constrained extremum

The utility of this observation is that the exact form of the surface Ω near x_0 is replaced by the simple description of its tangent in the expression for the necessary conditions. In order for the procedure to work, however, it must be possible to express the tangent in terms of the derivatives of the constraints. For this reason we introduce the definition of a regular point.

Definition. A point x_0 satisfying the constraints (1) is said to be a *regular point* of these constraints if the n linear functionals $g_1'(x_0)$, $g_2'(x_0)$, ..., $g_n'(x_0)$ are linearly independent.

The following theorem gives the necessary conditions satisfied by the solution of the constrained extremum problem.

Theorem 1. *If x_0 is an extremum of the functional f subject to the constraints $g_i(x) = 0$, $i = 1, 2, \ldots, n$; and if x_0 is a regular point of these constraints, then*

$$\delta f(x_0; h) = 0$$

for all h satisfying $\delta g_i(x_0; h) = 0$, $i = 1, 2, \ldots, n$.

Proof. Choose $h \in X$ such that $\delta g_i(x_0; h) = 0$ for $i = 1, 2, \ldots, n$. Let $y_1, y_2, \ldots, y_n \in X$ be n linearly independent vectors chosen so that the $n \times n$ matrix

$$M = \begin{bmatrix} \delta g_1(x_0; y_1) & \delta g_1(x_0; y_2) & \cdots & \delta g_1(x_0; y_n) \\ \delta g_2(x_0; y_1) & \delta g_2(x_0; y_2) & \cdots & \delta g_2(x_0; y_n) \\ \vdots & & & \\ \delta g_n(x_0; y_1) & & \cdots & \delta g_n(x_0; y_n) \end{bmatrix}$$

is nonsingular. The existence of such a set of vectors y_i follows directly from the regularity of the point x_0 (see Problem 11).

We now introduce the $n + 1$ real variables $\varepsilon, \varphi_1, \varphi_2, \ldots, \varphi_n$ and consider the n equations

$$\begin{aligned} &g_1(x_0 + \varepsilon h + \varphi_1 y_1 + \varphi_2 y_2 + \cdots + \varphi_n y_n) = 0 \\ &g_2(x_0 + \varepsilon h + \varphi_1 y_1 + \varphi_2 y_2 + \cdots + \varphi_n y_n) = 0 \\ &\vdots \\ &g_n(x_0 + \varepsilon h + \varphi_1 y_1 + \varphi_2 y_2 + \cdots + \varphi_n y_n) = 0. \end{aligned} \tag{2}$$

The Jacobian of this set with respect to the variables φ_i, at $\varepsilon = 0$ and $\varphi_i = 0$, is just the determinant of M and is therefore nonzero by assumption. Hence, the implicit function theorem (see, for example, Apostol [10] and also Section 9.2) applies and guarantees the existence of n functions $\varphi_i(\varepsilon)$ satisfying (2) and defined in some neighborhood of $\varepsilon = 0$.

Denote by $y(\varepsilon)$ the vector $\sum_{i=1}^{n} \varphi_i(\varepsilon) y_i$ and by $\varphi(\varepsilon)$ the n-dimensional vector having components $\varphi_i(\varepsilon)$. For each i we have

$$(3) \qquad 0 = g_i\left(x_0 + \varepsilon h + \sum_j \varphi_j y_j\right)$$
$$= g_i(x_0) + \varepsilon \delta g_i(x_0; h) + \delta g_i(x_0; y(\varepsilon)) + o(\varepsilon) + o[\|y(\varepsilon)\|].$$

Or, writing all n equations simultaneously and taking into account the fact that the first two terms on the right side of (3) are zero, we have, after taking the norm of the result,

$$(4) \qquad 0 = \|M\varphi(\varepsilon)\| + o(\varepsilon) + o[\|y(\varepsilon)\|].$$

Since M is nonsingular, there are constants $c_1 > 0$, $c_2 > 0$ such that $c_1\|\varphi(\varepsilon)\| \leq \|M\varphi(\varepsilon)\| \leq c_2\|\varphi(\varepsilon)\|$; and since the y_i's are linearly independent, there are constants $d_1 > 0$, $d_2 > 0$ such that $d_1\|y(\varepsilon)\| \leq \|\varphi(\varepsilon)\| \leq d_2\|y(\varepsilon)\|$. Hence, $c_1 d_1\|y(\varepsilon)\| \leq \|M\varphi(\varepsilon)\| \leq c_2 d_2\|y(\varepsilon)\|$ and, therefore (4) implies $\|y(\varepsilon)\| = o(\varepsilon)$. Geometrically, this result states that if one moves along the tangent plane of the constraint surface Ω from x_0 to $x_0 + \varepsilon h$, it is possible to get back to Ω by moving to $x_0 + \varepsilon h + y(\varepsilon)$, where $y(\varepsilon)$ is small compared with εh.

The points $x_0 + \varepsilon h + y(\varepsilon)$ define a one-parameter family of admissible vectors. Considering the functional f at these points, we must have

$$\frac{d}{d\varepsilon} f[x_0 + \varepsilon h + y(\varepsilon)]\bigg|_{\varepsilon=0} = 0.$$

Thus, since $\|y(\varepsilon)\| = o(\varepsilon)$, $\delta f(x_0; h) = 0$. ∎

From Theorem 1 it is easy to derive a finite-dimensional version of the Lagrange multiplier rule by using the following lemma.

Lemma 1. *Let $f_0, f_1, \ldots, f_n$ be linear functionals on a vector space X and suppose that $f_0(x) = 0$ for every $x \in X$ satisfying $f_i(x) = 0$ for $i = 1, 2, \ldots, n$. Then there are constants $\lambda_1, \lambda_2, \ldots, \lambda_n$ such that*

$$f_0 + \lambda_1 f_1 + \lambda_2 f_2 + \cdots + \lambda_n f_n = 0.$$

Proof. See Problem 16, Chapter 5 (which is solved by use of the Hahn-Banach theorem). ∎

Theorem 2. *If x_0 is an extremum of the functional f subject to the constraints*

$$g_i(x) = 0, \qquad i = 1, 2, \ldots, n,$$

and x_0 is a regular point of these constraints, then there are n scalar λ_i, $i = 1, 2, \ldots, n$, that render the functional

$$f(x) + \sum_{i=1}^{n} \lambda_i g_i(x)$$

stationary at x_0.

Proof. By Theorem 1 the differential $\delta f(x_0; h)$ is zero whenever ea of the differentials $\delta g_i(x_0; h)$ is zero. The result then follows immediately from Lemma 1. ∎

Example 1. Constrained problems of the form above are called isope metric problems in the calculus of variations since they were originally studied in connection with finding curves of given perimeter which ma: mized some objective such as enclosed area. Here we seek the curve in t $t-x$ plane having end points $(-1, 0)$, $(1, 0)$; length l; and enclosing maximum area between itself and the t-axis. Thus we wish to maximize[1]

$$\int_{-1}^{1} x(t)\, dt$$

subject to

$$\int_{-1}^{1} \sqrt{\dot{x}^2 + 1}\, dt = l.$$

Therefore, we seek a stationary point of

$$J(x) = \int_{-1}^{1} (x + \lambda\sqrt{\dot{x}^2 + 1})\, dt = \int_{-1}^{1} f(x, \dot{x}, t)\, dt.$$

Applying the Euler equations, we require

$$f_x - \frac{df_{\dot{x}}}{dt} = 0$$

or

$$1 - \lambda \frac{d}{dt} \frac{\dot{x}}{\sqrt{\dot{x}^2 + 1}} = 0$$

or

$$\frac{\dot{x}}{\sqrt{\dot{x}^2 + 1}} = \frac{1}{\lambda} t + c.$$

[1] See Problem 12 for a different formulation of this problem.

It is easily verified that a solution to this first-order differential equation is given by the arc of a circle

$$(x - x_1)^2 + (t - t_1)^2 = r^2.$$

The parameters x_1, t_1, and r are chosen to satisfy the boundary conditions and the condition on total length.

GLOBAL THEORY

7.8 Convex and Concave Functionals

Definition. A real-valued functional f defined on a convex subset C of a linear vector space is said to be *convex* if

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

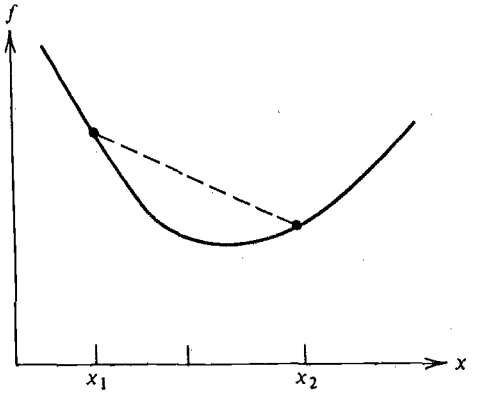

Figure 7.6 A convex function

for all $x_1, x_2 \in C$ and all α, $0 < \alpha < 1$. If strict inequality holds whenever $x_1 \neq x_2$, f is said to be *strictly convex*. A functional g defined on a convex set is said to be (*strictly*) *concave* if $-g$ is (strictly) convex.

Examples of convex functions in one dimension are $f(x) = x^2$; $f(x) = e^x$ for $x > 0$; and the discontinuous function

$$f(x) = \begin{cases} 1 & x = 0 \\ x^2 & x > 0 \end{cases}$$

defined on $[0, \infty)$. The functional

$$f(x) = \int_0^1 \{x^2(t) + |x(t)|\}\, dt$$

defined on $L_2[0, 1]$ is convex and continuous (the reader may wish to verify this). Any norm is a convex functional.

A convex functional defined on an infinite-dimensional normed space may be discontinuous everywhere since, for example, any linear functional is convex, and we constructed discontinuous linear functionals earlier.

Convex functionals play a special role in the theory of optimization because most of the theory of local extrema for general nonlinear functionals can be strengthened to become global theory when applied to convex functionals. Conversely, results derived for minimization of convex functionals often have analogs as local properties for more general problems. The study of convex functionals leads then not only to an aspect of optimization important in its own right but also to increased insight for a large portion of optimization theory.

The following proposition illustrates the global nature of results for minimization problems involving convex functionals.

Proposition 1. *Let f be a convex functional defined on a convex subset C of a normed space. Let $\mu = \inf_{x \in C} f(x)$. Then*

1. *The subset Ω of C where $f(x) = \mu$ is convex.*
2. *If x_0 is a local minimum of f, then $f(x_0) = \mu$ and, hence x_0 is a global minimum.*

Proof.

1. Suppose $x_1, x_2 \in \Omega$. Then for $x = \alpha x_1 + (1 - \alpha)x_2$, $0 < \alpha < 1$, we have $f(x) \leq \alpha f(x_1) + (1 - \alpha)f(x_2) = \mu$. But for any $x \in C$ necessarily $f(x) \geq \mu$. Thus, $f(x) = \mu$.
2. Suppose N is a neighborhood about x_0 in which x_0 minimizes f. For any $x_1 \in C$, there is an $x \in N$ such that $x = \alpha x_0 + (1 - \alpha)x_1$ for some α, $0 < \alpha < 1$. We have $f(x_0) \leq f(x) \leq \alpha f(x_0) + (1 - \alpha)f(x_1)$. Or $f(x_0) \leq f(x_1)$. ∎

The study of convex functionals is quickly and effectively reduced to the study of convex sets by considering the region above the graph of the functional.

Definition. In correspondence to a convex functional f defined on a convex set C in a vector space X, we define the convex set $[f, C]$ in $R \times X$ as

$$[f, C] = \{(r, x) \in R \times X : x \in C, f(x) \leq r\}.$$

Usually we think of the space $R \times X$ as being oriented so that the R axis, i.e., all vectors of the form (r, θ), is the vertical axis. Then the set $[f, C]$ can be thought of as the region above the graph of f, as illustrated in Figure 7.7. This set $[f, C]$ is sometimes called the *epigraph* of f over C.

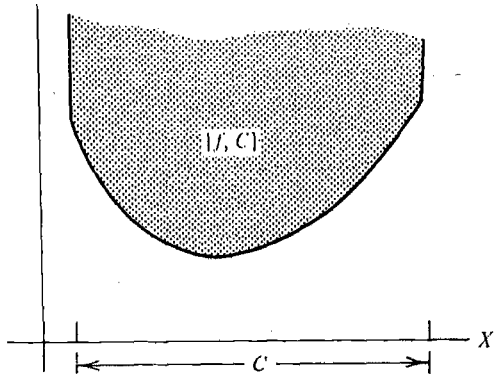

Figure 7.7 The convex region above the graph

Although we have no occasion to do so, we could imagine forming the set $[f, C]$ as in the above definition even if f were not convex. In any case, however, we have the following proposition.

Proposition 2. *The functional f defined on the convex domain C is convex if and only if $[f, C]$ is a convex set.*

The major portion of our analysis of convex functionals is based on consideration of the corresponding convex set of the last definition. To apply our arsenal of supporting hyperplane theorems to this set, however, it must be determined under what conditions the set $[f, C]$ contains interior points. The next section is devoted primarily to an analysis of this question. Since the results are so favorable—namely continuity of the functional at a single point guarantees an interior point—the reader need only glance through the proposition statements at first reading and proceed to the next section.

*7.9 Properties of the Set $[f, C]$

Proposition 1. *If f is a convex functional on the convex domain C in a normed space and C has nonempty relative interior $\mathring{C}$, then the convex set $[f, C]$ has a relative interior point (r_0, x_0) if and only if f is continuous at the point $x_0 \in \mathring{C}$.*

Proof. First assume that f is continuous at a point $x_0 \in \mathring{C}$. Denote by $N(x_0, \delta)$ the open spherical neighborhood of x_0 with radius δ. We note that $v([f, C])$, the linear variety generated by $[f, C]$, is equal to $R \times v(C)$. Given ε, $0 < \varepsilon < 1$, there is a $\delta > 0$ such that for $x \in N(x_0, \delta) \cap v(C)$ we have $x \in \mathring{C}$ and $|f(x) - f(x_0)| < \varepsilon$. Let $r_0 = f(x_0) + 2$. Then the point $(r_0, x_0) \in [f, C]$ is a relative interior point of $[f, C]$ since $(r, x) \in [f, C]$ for $|r - r_0| < 1$ and $x \in N(x_0, \delta) \cap v(C)$.

Now suppose that (r_0, x_0) is a relative interior point of $[f, C]$. Then there is $\varepsilon_0 > 0$, $\delta_0 > 0$ such that for $x \in N(x_0, \delta_0) \cap v(C)$ and $|r - r_0| < \varepsilon_0$ we have $r \geq f(x)$. Thus f is bounded above by $f(x_0) + \varepsilon_0$ on the neighborhood $N(x_0, \delta_0) \cap v(C)$.

We show now that the above implies that f is continuous at x_0. Without loss of generality we may assume $x_0 = \theta$ and $f(x_0) = 0$. For any ε, $0 < \varepsilon < 1$, and for any $x \in N(x_0, \varepsilon\delta_0) \cap v(C)$, we have

$$f(x) = f\left[(1 - \varepsilon)\theta + \varepsilon\left(\frac{1}{\varepsilon} x\right)\right] \leq (1 - \varepsilon) f(\theta) + \varepsilon f\left(\frac{1}{\varepsilon} x\right) \leq \varepsilon\varepsilon_0$$

where ε_0 is the bound on f in $N(x_0, \delta_0) \cap v(C)$. Furthermore,

$$0 = f(\theta) = f\left[\frac{1}{1 + \varepsilon} x + \left(1 - \frac{1}{1 + \varepsilon}\right)\left(-\frac{1}{\varepsilon} x\right)\right] \leq \frac{1}{1 + \varepsilon} f(x) + \left(1 - \frac{1}{1 + \varepsilon}\right) f\left(-\frac{1}{\varepsilon} x\right)$$

or

$$f(x) \geq -\varepsilon f\left(-\frac{1}{\varepsilon} x\right) \geq -\varepsilon\varepsilon_0.$$

Therefore, for $x \in N(x_0, \varepsilon\delta_0) \cap v(C)$, we have $|f(x)| \leq \varepsilon\varepsilon_0$. Thus f is continuous at x_0. ∎

Convex functionals enjoy many of the properties of linear functionals. As an example, the following proposition is a generalization of Proposition 1, Section 5.2.

Proposition 2. *A convex functional f defined on a convex domain C and continuous at a single point in the relative interior $\mathring{C}$ of C is continuous throughout $\mathring{C}$.*

Proof. Without loss of generality we may assume that f is continuous at $\theta \in \mathring{C}$ and that $f(\theta) = 0$. Furthermore, by restricting attention to $v(C)$, we may assume that C has interior points rather than relative interior points.

Let y be an arbitrary point in $\mathring{C}$. Since $\mathring{C}$ is (relatively) open, there is a $\beta > 1$ such that $\beta y \in \mathring{C}$. Given $\varepsilon > 0$, let $\delta > 0$ be such that $\|x\| < \delta$ implies $|f(x)| < \varepsilon$. Then for $\|z - y\| < (1 - \beta^{-1})\delta$, we have

$$z = y + (1 - \beta^{-1})x = \beta^{-1}(\beta y) + (1 - \beta^{-1})x$$

for some $x \in \mathring{C}$ with $\|x\| < \delta$. Thus $z \in C$ and

$$f(z) \leq \beta^{-1} f(\beta y) + (1 - \beta^{-1}) f(x) < \beta^{-1} f(\beta y) + (1 - \beta^{-1})\varepsilon.$$

Thus f is bounded above in the sphere $\|z - y\| < (1 - \beta^{-1})\delta$. It follows that for sufficiently large r the point (r, y) is an interior point of $[f, C]$; hence, by Proposition 1, f is continuous at y. ∎

The proof of the following important corollary is left to the reader.

Corollary 1. *A convex functional defined on a finite-dimensional convex set C is continuous throughout $\mathring{C}$.*

Having established the simple relation between continuity and interior points, we conclude this section by noting a property of f which holds *if* $[f, C]$ happens to be closed. As illustrated in Figure 7.8, closure of $[f, C]$ is related to the continuity properties of f on the boundary of C.

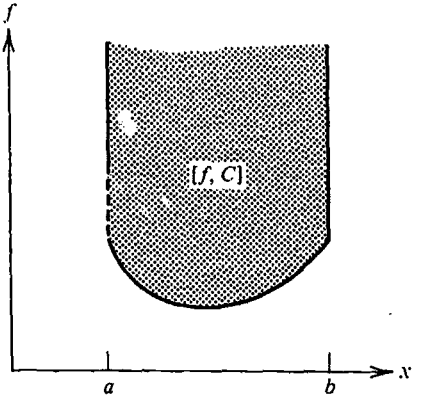

Figure 7.8 A nonclosed epigraph

Proposition 3. *If $[f, C]$ is closed, then f is lower semicontinuous on C.*

Proof. The set $\{(a, x) \in R \times X : x \in X\}$ is obviously closed for each $a \in R$. Hence, if $[f, C]$ is closed, so is

$$[f, C] \cap \{(a, x) : x \in X\} = \{(a, x) : x \in C, f(x) \leq a\}$$

for each $a \in R$. It follows that the set

$$T_a = \{x : x \in C, f(x) \leq a\}$$

is closed.

Now suppose $\{x_i\}$ is a sequence from C converging to $x \in C$. Let $b = \liminf_{x_i \to x} f(x_i)$. If $b = -\infty$, then $x \in \overline{T}_a = T_a$ for each $a \in R$ which is impossible. Thus $b > -\infty$ and $x \in \overline{T}_{b+\varepsilon} = T_{b+\varepsilon}$ for all $\varepsilon > 0$. In other words, $f(x) \leq \liminf_{x_i \to x} f(x_i)$ which proves that f is lower semicontinuous. ∎

Figure 7.9 shows the graph of a convex functional f defined on a disk C in E^2 that has closed $[f, C]$ but is discontinuous (although lower semicontinuous) at a point x.

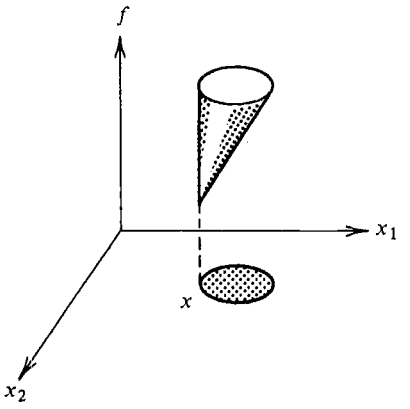

Figure 7.9

7.10 Conjugate Convex Functionals

A purely abstract approach to the theory of convex functionals, including a study of the convex set $[f, C]$ as in the previous section, leads quite naturally to an investigation of the dual representation of this set in terms of closed hyperplanes. The concept of conjugate functionals plays a natural and fundamental role in such a study. As an important consequence of this investigation, we obtain a very general duality principle for optimization problems which extends the earlier duality results for minimum norm problems.

Definition. Let f be a convex functional defined on a convex set C in a normed space X. The *conjugate set* C^* is defined as

$$C^* = \{x^* \in X^* : \sup_{x \in C} [\langle x, x^* \rangle - f(x)] < \infty\}$$

and the *functional f^* conjugate to f* is defined on C^* as

$$f^*(x^*) = \sup_{x \in C} [\langle x, x^* \rangle - f(x)].$$

Proposition 1. *The conjugate set C^* and the conjugate functional f^* are convex and $[f^*, C^*]$ is a closed convex subset of $R \times X^*$.*

Proof. For any $x_1^*, x_2^* \in X^*$ and any α, $0 < \alpha < 1$, we have

$$\begin{aligned}\sup_{x \in C} \{\langle x, \alpha x_1^* + (1-\alpha)x_2^* \rangle - f(x)\} &= \sup_{x \in C} \{\alpha[\langle x, x_1^* \rangle - f(x)] \\ &\quad + (1-\alpha)[\langle x, x_2^* \rangle - f(x)]\} \\ &\leq \alpha \sup_{x \in C} [\langle x, x_1^* \rangle - f(x)] \\ &\quad + (1-\alpha) \sup_{x \in C} [\langle x, x_2^* \rangle - f(x)]\end{aligned}$$

from which it follows immediately that C^* and f^* are convex.

Next we prove that $[f^*, C^*]$ is closed. Let $\{(s_i, x_i^*)\}$ be a convergent sequence from $[f^*, C^*]$ with $(s_i, x_i^*) \to (s, x^*)$. We show now that $(s, x^*) \in [f^*, C^*]$. For every i and every $x \in C$, we have

$$s_i \geq f^*(x_i^*) \geq \langle x, x_i^* \rangle - f(x).$$

Taking the limit as $i \to \infty$, we obtain

$$s \geq \langle x, x^* \rangle - f(x)$$

for all $x \in C$. Therefore,

$$s \geq \sup_{x \in C} [\langle x, x^* \rangle - f(x)]$$

from which it follows that $x^* \in C^*$ and $s \geq f^*(x^*)$. ∎

We see that the conjugate functional defines a set $[f^*, C^*]$ which is of the same type as $[f, C]$; therefore we write $[f, C]^* = [f^*, C^*]$. Note that if $f = 0$, the conjugate functional f^* becomes the support functional of C.

Example 1. Let $X = C = E^n$ and define, for $x = (x_1, x_2, \ldots, x_n)$, $f(x) = 1/p \sum_{i=1}^n |x_i|^p$, $1 < p < \infty$. Then for $x^* = (\xi_1, \xi_2, \ldots, \xi_n)$,

$$f^*(x^*) = \sup \left[\langle x, x^* \rangle - \frac{1}{p} \sum_{i=1}^n |x_i|^p \right] = \sup \left[\sum_{i=1}^n \xi_i x_i - \frac{1}{p} \sum_{i=1}^n |x_i|^p \right].$$

The supremum on the right is achieved by some x since the problem is finite dimensional. We find, by differentiation, the solution

$$\xi_i = |x_i|^{p-1} \operatorname{sgn} x_i$$

$$f^*(x^*) = \sum_{i=1}^{n} |x_i|^p \left(1 - \frac{1}{p}\right) = \frac{1}{q} \sum_{i=1}^{n} |\xi_i|^q$$

where $1/p + 1/q = 1$.

Let us investigate the relation of the conjugate functional to separating hyperplanes. On the space $R \times X$, closed hyperplanes are represented by an equation of the form

$$sr + \langle x, x^* \rangle = k$$

where s, k, and x^* determine the hyperplane. Recalling that we agreed to refer to the R axis as vertical, we say that a hyperplane is nonvertical if it intersects the R axis at one and only one point. This is equivalent to the requirement that the defining linear functional (s, x^*) have $s \neq 0$. If attention is restricted to nonvertical hyperplanes, we may, without loss of generality, consider only those linear functionals of the form $(-1, x^*)$. Any nonvertical closed hyperplane can then be obtained by appropriate choice of x^* and k.

To develop a geometric interpretation of the conjugate functional, note that as k varies, the solutions (r, x) of the equation

$$\langle x, x^* \rangle - r = k$$

describe parallel closed hyperplanes in $R \times X$. The number $f^*(x^*)$ is the supremum of the values of k for which the hyperplane intersects $[f, C]$. Thus the hyperplane $\langle x, x^* \rangle - r = f^*(x^*)$ is a support hyperplane of $[f, C]$.

In the terminology of Section 5.13, $f^*(x^*)$ is the support functional $h[(-1, x^*)]$ of the functional $(-1, x^*)$ for the convex set $[f, C]$. The special feature here is that we only consider functionals of the form $(-1, x^*)$ on $R \times X$ and thereby eliminate the need of carrying an extra variable.

For the application to optimization problems, the most important geometric interpretation of the conjugate functional is that it measures vertical distance to the support hyperplane. The hyperplane

$$\langle x, x^* \rangle - r = f^*(x^*)$$

intersects the vertical axis (i.e., $x = \theta$) at $(-f^*(x^*), \theta)$. Thus, $-f^*(x^*)$ is the vertical height of the hyperplane above the origin. (See Figure 7.10.)

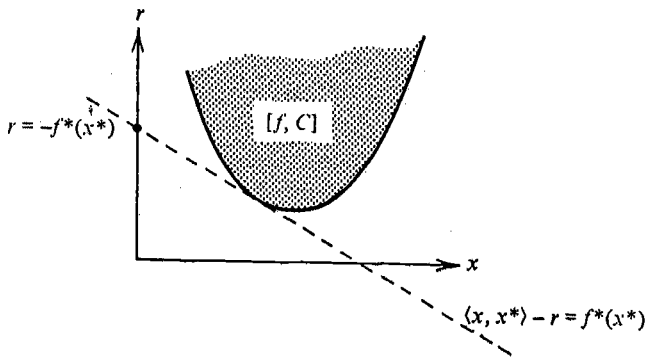

Figure 7.1C A conjugate convex functional

Another interpretation more clearly illuminates the duality between $[f, C]$ and $[f^*, C^*]$ in terms of the dual representation of a convex set as a collection of points or as the intersection of half-spaces. Given the point $(s, x^*) \in R \times X^*$, let us associate the half-space consisting of all $(r, x) \in R \times X$ satisfying

$$\langle x, x^* \rangle - r \leq s.$$

Then the set $[f^*, C^*]$ consists of those (nonvertical) half-spaces that contain the set $[f, C]$. Hence $[f^*, C^*]$ is the dual representation of $[f, C]$.

Beginning with an arbitrary convex functional φ defined on a convex subset Γ of a dual space X^*, we may, of course, define the conjugate of φ in X^{**} or, alternatively, following the standard pattern for duality relations (e.g., see Section 5.7), define the set ${}^*\Gamma$ in X as

$${}^*\Gamma = \{x : \sup_{x^* \in \Gamma} [\langle x, x^* \rangle - \varphi(x^*)] < \infty\}$$

and the convex functional

$${}^*\varphi(x) = \sup_{x^* \in \Gamma} [\langle x, x^* \rangle - \varphi(x^*)]$$

on ${}^*\Gamma$. We then write ${}^*[\varphi, \Gamma] = [{}^*\varphi, {}^*\Gamma]$. With these definitions we have the following characterization of the duality between a convex functional and its conjugate.

Proposition 2. *Let f be a convex functional on the convex set C in a normed space X. If $[f, C]$ is closed, then $[f, C] = {}^*[[f, C]^*]$.*

Proof. We show first that $[f, C] \subset {}^*[f^*, C^*] = {}^*[[f, C]^*]$. Let $(r, x) \in [f, C]$; then for all $x^* \in C^*$, $f^*(x^*) \geq \langle x, x^* \rangle - f(x)$. Hence, we have $r \geq f(x) \geq \langle x, x^* \rangle - f^*(x^*)$ for all $x^* \in C^*$. Thus

$$r \geq \sup_{x^* \in C^*} [\langle x, x^* \rangle - f^*(x^*)]$$

and $(r, x) \in {}^*[f^*, C^*]$.

We prove the converse by contraposition. Let $(r_0, x_0) \notin [f, C]$. Since $[f, C]$ is closed, there is a hyperplane separating (r_0, x_0) and $[f, C]$. Thus there exist $x^* \in X^*$, s, and c such that

$$sr + \langle x, x^* \rangle \leq c < sr_0 + \langle x_0, x^* \rangle$$

for all $(r, x) \in [f, C]$. It can be shown that, without loss of generality, this hyperplane can be assumed to be nonvertical and hence $s \neq 0$ (see Problem 16). Furthermore, since r can be made arbitrarily large, we must have $s < 0$. Thus we take $s = -1$. Now it follows that $\langle x, x^* \rangle - f(x) \leq c$ for all $x \in C$, which implies that $(c, x^*) \in [f^*, C^*]$. On the other hand, $c < \langle x_0, x^* \rangle - r_0$ implies $\langle x_0, x^* \rangle - c > r_0$, which implies that $(r_0, x_0) \notin {}^*[f^*, C^*]$. ∎

7.11 Conjugate Concave Functionals

A development similar to that of the last section applies to concave functionals. It must be stressed, however, that we *do not* treat concave functionals by merely multiplying by -1 and then applying the theory for convex functionals. There is an additional sign change in part of the definition. See Problem 15.

Given a concave functional g defined on a convex subset D of a vector space, we define the set

$$[g, D] = \{(r, x) : x \in D, r \leq g(x)\}.$$

The set $[g, D]$ is convex and all of the results on continuity, interior points, etc., of Section 7.9 have direct extensions here.

Definition. Let g be a concave functional on the convex set D. The *conjugate set* D^* is defined as

$$D^* = \{x^* \in X^* : \inf_{x \in D} [\langle x, x^* \rangle - g(x)] > -\infty\},$$

and the *functional* g^* *conjugate to* g is defined as

$$g^*(x^*) = \inf_{x \in D} [\langle x, x^* \rangle - g(x)].$$

We can readily verify that D^* is convex and that g^* is concave. We write $[g, D]^* = [g^*, D^*]$.

Since our notation does not completely distinguish between the development for convex and concave functionals, it is important to make clear which is being employed in any given context. This is particularly true when the original function is linear, since either definition of the conjugate functional might be employed and, in general, they are not equal.

The geometric interpretation for concave conjugate functionals is similar to that for convex conjugate functionals. The hyperplane $\langle x, x^* \rangle - r = g^*(x^*)$ supports the set $[g, D]$. Furthermore, $-g^*(x^*)$ is the intercept of that hyperplane with the vertical axis. The situation is summarized in Figure 7.11.

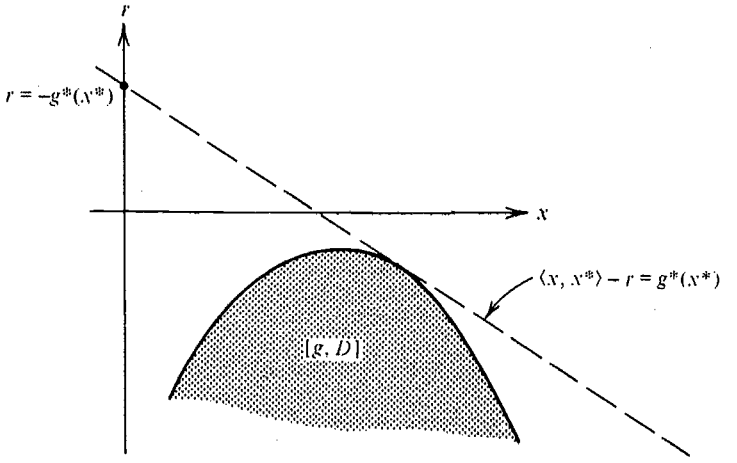

Figure 7.11 A conjugate concave functional

7.12 Dual Optimization Problems

We come now to the application of conjugate functionals to optimization. Suppose we seek to minimize a convex functional over a convex domain. Or more generally, if f is convex over C and g is concave over D, suppose we seek

$$\inf_{C \cap D} [f(x) - g(x)].$$

In standard minimization problems, g is usually zero. But as we shall see, the present generalization is conceptually helpful. The general problem is illustrated in Figure 7.12. The problem can be interpreted as that of finding the minimum vertical separation of the sets $[f, C]$ and $[g, D]$. It is reasonably clear, from geometric intuition, that this distance is equal to the maximum vertical separation of two parallel hyperplanes separating $[f, C]$ and $[g, D]$. This relation, between a given minimization problem and an equivalent maximization problem, is a generalization of the duality principle for minimum norm problems.

Conjugate functionals are precisely what is needed for expressing this duality principle algebraically. Since $-f^*(x^*)$ is the vertical distance to a

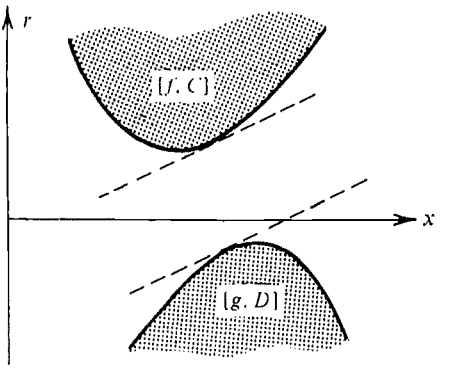

Figure 7.12

support hyperplane below $[f, C]$, and $-g^*(x^*)$ is the vertical distance to the parallel support hyperplane above $[g, D]$, $g^*(x^*) - f^*(x^*)$ is the vertical separation of the two hyperplanes. The duality principle is stated in detail in the following theorem.

Theorem 1. (*Fenchel Duality Theorem*) *Assume that f and g are, respectively, convex and concave functionals on the convex sets C and D in a normed space X. Assume that $C \cap D$ contains points in the relative interior of C and D and that either $[f, C]$ or $[g, D]$ has nonempty interior. Suppose further that $\mu = \inf_{x \in C \cap D} \{f(x) - g(x)\}$ is finite. Then*

$$\mu = \inf_{x \in C \cap D} \{f(x) - g(x)\} = \max_{x^* \in C^* \cap D^*} \{g^*(x^*) - f^*(x^*)\}$$

where the maximum on the right is achieved by some $x_0^* \in C^* \cap D^*$.

If the infimum on the left is achieved by some $x_0 \in C \cap D$, then

$$\max_{x \in C} [\langle x, x_0^* \rangle - f(x)] = \langle x_0, x_0^* \rangle - f(x_0)$$

and

$$\min_{x \in D} [\langle x, x_0^* \rangle - g(x)] = \langle x_0, x_0^* \rangle - g(x_0).$$

Proof. By definition, for all $x^* \in C^* \cap D^*$, $x \in C \cap D$,

$$f^*(x^*) \geq \langle x, x^* \rangle - f(x)$$

$$g^*(x^*) \leq \langle x, x^* \rangle - g(x).$$

Thus,

$$f(x) - g(x) \geq g^*(x^*) - f^*(x^*)$$

and hence

$$\inf_{C\cap D}[f(x)-g(x)] \geq \sup_{C^*\cap D^*}[g^*(x^*)-f^*(x^*)].$$

Therefore, the equality in the theorem is proved if an $x_0^* \in C^* \cap D^*$ can be found for which $\inf_{C\cap D}[f(x)-g(x)] = g^*(x_0^*)-f^*(x_0^*)$.

The convex set $[f-\mu, C]$ is a vertical displacement of $[f, C]$; by definition of μ the sets $[f-\mu, C]$ and $[g, D]$ are arbitrarily close but have disjoint relative interiors. Therefore, since one of these sets has nonempty interior, there is a closed hyperplane in $R \times X$ separating them. This hyperplane cannot be vertical because otherwise its vertical projection onto X would separate C and D. Since the hyperplane is not vertical, it can be represented as $\{(r, x) \in R \times X: \langle x, x_0^*\rangle - r = c\}$ for some $x_0^* \in X^*$ and $c \in R$. Now since $[g, D]$ lies below this hyperplane but is arbitrarily close to it, we have

$$c = \inf_{x\in D}[\langle x, x_0^*\rangle - g(x)] = g^*(x_0^*).$$

Likewise,

$$c = \sup_{x\in C}[\langle x, x_0^*\rangle - f(x)+\mu] = f^*(x_0^*)+\mu.$$

Thus $\mu = g^*(x_0^*)-f^*(x_0^*)$.

If the infimum μ on the left is achieved by some $x_0 \in C \cap D$, the sets $[f-\mu, C]$ and $[g, D]$ have the point $(g(x_0), x_0)$ in common and this point lies in the separating hyperplane. ∎

In typical applications of this theorem, we consider minimizing a convex functional f on a convex domain D; the set D representing constraints. Accordingly, we take $C = X$ and $g = 0$ in the theorem. Calculation of f^* is itself an optimization problem, but with $C = X$ it is unconstrained. Calculation of g^* is an optimization problem with a linear objective functional when $g = 0$.

Example 1. (An Allocation Problem) Suppose that there is a fixed quantity x_0 of some commodity (such as money) which is to be allocated among n distinct activities in such a way as to maximize the total return from these activities. We assume that the return associated with the i-th activity when allocated x_i units is $g_i(x_i)$ where g_i, due to diminishing marginal returns, is assumed to be an increasing concave function. In these terms the problem is one of finding $x = (x_1, x_2, \ldots, x_n)$ so as to

$$(1)\qquad \begin{cases} \text{maximize } g(x) = \sum_{i=1}^{n} g_i(x_i) \\ \text{subject to } \sum_{i=1}^{n} x_i = x_0, \qquad x_i \geq 0 \qquad i = 1, 2, \ldots, n. \end{cases}$$

To solve this problem by conjugate functionals, we set D equal to the positive orthant, f equal to the zero functional, and $C = \{x : \sum_{i=1}^{n} x_i = x_0\}$. The functional conjugate to f is (for $y = (y_1, y_2, \ldots, y_n)$)

$$f^*(y) = \sup_{\Sigma x_i = x_0} y'x.$$

This is finite only if $y = \lambda(1, 1, \ldots, 1)$ in which case it is equal to λx_0. Thus

$$C^* = \{y : y = \lambda(1, 1, \ldots, 1)\}$$
$$f^*(\lambda(1, 1, \ldots, 1, 1)) = \lambda x_0.$$

Also, for each i we define the conjugate functions (of a single variable)

$$g_i^*(y_i) = \inf_{x_i \geq 0} [x_i y_i - g_i(x_i)] \tag{2}$$

and it is clear that

$$g^*(y) = \sum_{i=1}^{n} g_i^*(y_i).$$

The problem conjugate to (1) is therefore

$$\min_{\lambda} \left[\lambda x_0 - \sum_{i=1}^{n} g_i^*(\lambda) \right]. \tag{3}$$

In this form we note that to solve the allocation problem requires evaluation of the conjugate functions g_i^* and then solution of (3) which is minimization with respect to the single variable λ. Once the optimal value of λ is determined the x_i's which solve (1) can be found to be those which minimize (2) with each $y_i = \lambda$.

This analysis can be modified in order to apply to a multistage allocation problem where there is the possibility of investment growth of uncommitted resources. See Problem 19.

Example 2. (Horse-Racing Problem) What is the best way to place bets totaling x_0 dollars in a race involving n horses? Assume we know p_i, the probability that the i-th horse wins, and s_i, the amount that the rest of the public is betting on the i-th horse. The track keeps a proportion $1 - C$ of the total amount bet ($0 < 1 - C < 1$) and distributes the rest among the public in proportion to the amounts bet on the winning horse.

Symbolically, if we bet x_i on the i-th horse, $i = 1, 2, \ldots, n$, we receive

$$C\left(x_0 + \sum_{i=1}^{n} s_i\right) \frac{x_i}{s_i + x_i}$$

if the i-th horse wins. Thus the expected net return, R, is

$$R = C\left(x_0 + \sum_{i=1}^{n} s_i\right)\left\{\sum_{i=1}^{n} \frac{p_i x_i}{s_i + x_i}\right\} - x_0. \tag{4}$$

Our problem is to find x_i, $i = 1, 2, \ldots, n$, which maximize R subject to

$$\sum_{i=1}^{n} x_i = x_0, \qquad x_i \geq 0, \qquad i = 1, 2, \ldots, n \tag{5}$$

or, equivalently, to maximize

$$\sum_{i=1}^{n} g_i(x_i) \tag{6}$$

subject to (5), where

$$g_i(x_i) = \frac{p_i x_i}{s_i + x_i}.$$

This problem is exactly the type treated in Example 1 since each g_i is concave for positive x_i (as can be verified by observing that the second derivative is negative). Solution to the problem is obtained by calculating the functions conjugate to the g_i's.

A typical g_i is shown in Figure 7.13. Its slope at $x_i = 0$ is p_i/s_i and it approaches the value p_i as $x_i \to \infty$. The value of the conjugate functional g_i^* at λ is obtained by finding the lowest line of slope λ which lies above g_i. It is clear from the diagram that for $\lambda \geq p_i/s_i$, we have $g_i^*(\lambda) = 0$, and that for $\lambda < 0$, $g_i^*(\lambda)$ is not defined. For $0 < \lambda < p_i/s_i$, we have

$$g_i^*(\lambda) = \min_{x_i > 0} [\lambda x_i - g_i(x_i)]. \tag{7}$$

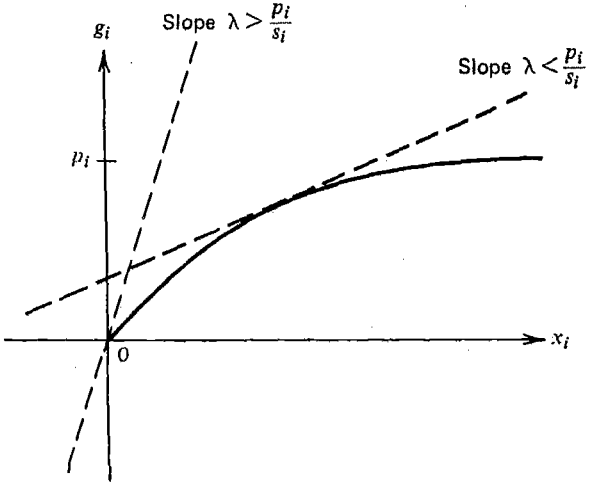

Figure 7.13

Performing the minimization by differentiation, we have the equation

$$\lambda = \frac{s_i p_i}{(s_i + x_i)^2}$$

or

$$x_i = \sqrt{\frac{s_i p_i}{\lambda}} - s_i. \tag{8}$$

Substitution back into (7) yields

$$g_i^*(\lambda) = \begin{cases} \dfrac{-p_i x_i^2}{(s_i + x_i)^2} & \text{for } 0 < \lambda < \dfrac{p_i}{s_i} \\ 0 & \text{for } \lambda \geq \dfrac{p_i}{s_i} \end{cases} \tag{9}$$

where x_i is determined from equation (8).

We can now deduce the form of the answer. Suppose that λ has been found from (3). To simplify the notation, rearrange the indices so that $p_1/s_1 > p_2/s_2 > \cdots > p_n/s_n$ (assuming strict inequality). For the given λ, we define m as the largest index for which $p_i/s_i \geq \lambda$. Then, from equations (8) and (9), our solution may be written

$$x_i = \begin{cases} \sqrt{\dfrac{s_i p_i}{\lambda}} - s_i & \text{for } i = 1, 2, \ldots, m \\ 0 & \text{for } i = m+1, \ldots, n. \end{cases} \tag{10}$$

The parameter λ in this section can be found from (3) or from the constraint (5). In other words, λ is chosen so that

$$S(\lambda) = \sum_{p_i/s_i \geq \lambda} \left(\sqrt{\frac{s_i p_i}{\lambda}} \right) - s_i = x_0. \tag{11}$$

Now $S(\lambda)$ is easily verified to be continuous and $S(0) = \infty$, $S(\infty) = 0$. Thus there is a λ satisfying (11).

Note that for small x_0 ($x_0 << \max s_i$), the total amount should be bet on a single horse, the horse corresponding to the maximum p_i/s_i, or equivalently, the maximum $p_i r_i$ where $r_i = C \sum_j s_j/s_i$ is the track odds.

Example 3. Consider the minimum energy control problem discussed in Sections 3.10 and 6.10. We seek the element $u \in L_2[0, 1]$ minimizing

$$f(u) = \tfrac{1}{2} \int_0^1 u^2(t)\, dt,$$

while satisfying the linear constraints

$$Ku = c$$

where $K: L_2[0,1] \to E^n$.

In Theorem 1, let $C = L_2[0, 1]$, $g = 0$, and $D = \{u : Ku = c\}$. We assume that D is nonempty. Then we can readily verify that

$$C^* = L_2[0, 1]$$

and

$$f^*(u^*) = \tfrac{1}{2} \int_0^1 [u^*(t)]^2 \, dt.$$

Since $g = 0$, the conjugate functional of the set D is equal to the support functional. Thus

$$D^* = \{u^* : u^* = K^*a, a \in E^n\}$$

and $g^*(K^*a) = (a \mid c)$.

The dual problem is therefore the finite-dimensional problem in the vector a:

$$\max \{(a \mid c) - \tfrac{1}{2}(K^*a \mid K^*a)\}$$

which is solved by finding the n vector a satisfying

$$KK^*a = c$$

where KK^* is an $n \times n$ matrix.

Finally, the solution to the original problem can be found in terms of the solution of the dual by application of the second part of Theorem 1. Thus

$$u_0 = K^*a.$$

*7.13 Min-Max Theorem of Game Theory

In this section we briefly introduce the classical theory of games and prove the fundamental min-max theorem. Our purpose is to show that the min-max theorem can be regarded as an example of the Fenchel duality theorem.

Let X be a normed space and X^* its normed dual. Let A be a fixed subset of X and B a fixed subset of X^*. In the form of game that we consider, one player (player A) selects a vector from his strategy set A while his opponent (player B) selects a vector x^* from his strategy set B. When both players have selected their respective vectors, the quantity

$\langle x, x^* \rangle$ is computed and player A pays that amount (in some appropriate units) to player B. Thus A seeks to make his selection so as to minimize $\langle x, x^* \rangle$ while B seeks to maximize $\langle x, x^* \rangle$.

Assuming for the moment that the quantities

$$\mu^0 = \min_{x \in A} \max_{x^* \in B} \langle x, x^* \rangle$$

$$\mu_0 = \max_{x^* \in B} \min_{x \in A} \langle x, x^* \rangle$$

exist, we first take the viewpoint of A in this game. By selecting $x \in A$, he loses no more than $\max_{x^* \in B} \langle x, x^* \rangle$; hence, by proper choice of x, say x_0, he can be assured of losing no more than μ^0. On the other hand, player B by selecting $x^* \in B$, wins at least $\min_{x \in A} \langle x, x^* \rangle$; therefore, by proper choice of x^*, say x_0^*, he can be assured of winning at least μ_0. It follows that $\mu_0 \leq \langle x_0, x_0^* \rangle \leq \mu^0$, and the fundamental question that arises is whether $\mu_0 = \mu^0$ so that there is determined a unique pay-off value for optimal play by both players.

The most interesting type of game that can be put into the form outlined above is the classical finite game. In a finite game each player has a finite set of strategies and the pay-off is determined by a matrix Q, the pay-off being q_{ij} if A uses strategy i and B uses strategy j. For example, in a simple coin-matching game, the players independently select either "heads" or "tails." If their choices match, A pays B 1 unit while, if they differ, B pays A 1 unit. The pay-off matrix in this case is

$$Q = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Finite games of this kind usually do not have a unique pay-off value. We might, however, consider a long sequence of such games and "randomized" strategies where each player determines his play in any one game according to a fixed probability distribution among his choices. Assuming that A has n basic strategies, he selects an n vector of probabilities $x = (x_1, x_2, \ldots, x_n)$ such that $x_i \geq 0$, $\sum_{i=1}^n x_i = 1$. Likewise, if B has m strategies, he selects $y = (y_1, y_2, \ldots, y_m)$ such that $y_i \geq 0$, $\sum_{i=1}^m y_i = 1$. The expected (or average) pay-off is then $(x \mid Qy)$.

Defining $A = \{x : x_i \geq 0, \sum_{i=1}^n x_i = 1\} \subset E^n$, and $B = \{x^* : x^* = Qy, y_i \geq 0, \sum_{i=1}^m y_i = 1\} \subset E^n$, the randomized game takes the standard form given at the beginning of the section. Note that A and B are both bounded closed convex sets. Other game-type optimization problems with bilinear objectives other than the classical randomized finite game also take this form.

We now give a simple proof of the min-max theorem based on duality. For simplicity, our proof is for reflexive spaces, although more general

versions of the result hold. For the generalizations, consult the references at the end of the chapter.

Theorem 1. (Min-Max) *Let X be a reflexive normed space and let A and B be compact convex subsets of X and X^*, respectively. Then*

$$\min_{x \in A} \max_{x^* \in B} \langle x, x^* \rangle = \max_{x^* \in B} \min_{x \in A} \langle x, x^* \rangle.$$

Proof. Define the functional f on X by

$$f(x) = \max_{x^* \in B} \langle x, x^* \rangle.$$

The maximum exists for each $x \in X$ since B is compact. The functional is easily shown to be convex and continuous on X. We seek an expression for

$$\min_{x \in A} f(x)$$

which exists by the compactness of A and the continuity of f. We now apply the Fenchel duality theorem with the associations: $f \to f$, $C \to X$, $g \to 0$, $D \to A$. We have immediately:

$$D^* = X^* \tag{1}$$

$$g^*(x^*) = \min_{x \in A} \langle x, x^* \rangle. \tag{2}$$

We claim that furthermore

$$C^* = B \tag{3}$$

$$f^*(x^*) = 0. \tag{4}$$

To prove (3) and (4), let $x_1^* \notin B$, and by using the separating hyperplane theorem, let $x_1 \in X$ and α be such that $\langle x_1, x_1^* \rangle - \langle x_1, x^* \rangle > \alpha > 0$ for all $x^* \in B$. Then $\langle x, x_1^* \rangle - \max_{x^* \in B} \langle x, x^* \rangle$ can be made arbitrarily large by taking $x = kx_1$ with $k > 0$. Thus

$$\sup_x [\langle x, x_1^* \rangle - f(x)] = \infty$$

and $x_1^* \notin C^*$.

Conversely, if $x_1^* \in B$, then $\langle x, x_1^* \rangle - \max_{x^* \in B} \langle x, x^* \rangle$ achieves a maximum value of 0 at $x = \theta$. This establishes (3) and (4).

The final result follows easily from the equality

$$\min_{x \in A} f(x) = \max_{x^* \in B \cap X^*} g^*(x^*) = \max_{x^* \in B} \min_{x \in A} \langle x, x^* \rangle. \quad \blacksquare$$

An interesting special case of the min-max theorem is obtained by taking B to be the unit sphere in X^*. In that case we obtain

$$\min_{x \in A} \|x\| = \max_{\|x^*\| \leq 1} - h(x^*)$$

where h is the support functional of the convex set A. This result is the duality theorem for minimum norm problems of Section 5.8.

7.14 Problems

1. On the vector space of continuous functions on $[0, 1]$, define

$$f(x) = \max_{0 \leq t \leq 1} x(t).$$

Determine for which x the Gateaux differential $\delta f(x; h)$ exists and is linear in h.

2. Repeat Problem 1 for

$$f(x) = \int_0^1 |x(t)|\, dt.$$

3. Show that the functional

$$f(x_1, x_2) = \begin{cases} \dfrac{x_1 x_2^{\,2}}{(x_1^{\,2} + x_2^{\,4})} & \text{if } x_1 \neq 0 \\ 0 & \text{if } x_1 = 0 \end{cases}$$

is Gateaux differentiable but not continuous at $x_1 = x_2 = 0$.

4. On the space $X = C[0, 1]$, define the functional $f(x) = [x(\frac{1}{2})]^2$. Find the Fréchet differential and Fréchet derivative of f.
5. Let φ be a function of a single variable having a continuous derivative and satisfying $|\varphi(\xi)| < K|\xi|$. Find the Gateaux differential of the functional $f(x) = \sum_{i=1}^{\infty} \varphi(\xi_i)$ where $x = \{\xi_i\} \in l_1$. Is this also a Fréchet differential?
6. Suppose the real-valued functional f defined on an open subset D of a normed space has a relative minimum at $x_0 \in D$. Show that if f is twice Gateaux differentiable at x_0, then $\langle h, f''(x_0)h \rangle \geq 0$ for all $h \in X$.
7. Let f be a real-valued functional defined on an open region D in a normed space X. Suppose that at $x_0 \in D$ the first Fréchet differential vanishes identically on X; within a sphere, $S(x_0, \varepsilon)$, $f''(x)$ exists and is continuous in x; and the lower bound of $\langle h, f''(x_0)h \rangle$ for $\|h\| = 1$ is positive. Show that f obtains a relative minimum at x_0.
8. Let A be a nonempty subset of a normed space X and let $x_0 \in A$. Denote by $C(A, x_0)$ the closed cone generated by $A - x_0$, i.e., the

intersection of all closed cones containing $A - x_0$. The *local closed cone* of A at x_0 is the set

$$LC(A, x_0) = \bigcap_{N \in \mathcal{N}} C(A \cap N, x_0)$$

where $\mathcal{N}$ is the class of all neighborhoods of x_0. Suppose that over A the Fréchet differentiable functional f achieves a minimum at x_0. Show that

$$f'(x_0) \in [LC(A, x_0)]^{\oplus}.$$

9. In some problems in the calculus of variations, it is necessary to consider a broader class of functions than usual. Suppose that we seek to extremize

$$J = \int_a^b F[x, \dot{x}, t]\, dt$$

among all functions x for which $x(a) = A$, $x(b) = B$ and which have continuous derivatives on $[a, b]$ except possibly at a single point in $[a, b]$. If the extremum has a continuous derivative except at $c \in [a, b]$, show that $F_x = dF_{\dot{x}}/dt$ on the intervals $[a, c)$ and $(c, b]$ and that the functions $F_{\dot{x}}$ and $F - \dot{x}F_{\dot{x}}$ are continuous at c. These are called the Weierstrass-Erdman corner conditions. Apply these considerations to the functional

$$J[x] = \int_{-1}^{1} x^2(1 - \dot{x})^2\, dt, \qquad x(-1) = 0, \qquad x(1) = 1.$$

10. Consider the effect of an additional source of constant income in the estate-planning problem.
11. Let $f_1, f_2, \ldots, f_n$ be linearly independent linear functionals on a vector space X. Show that there are n elements $x_1, x_2, \ldots, x_n$ in X such that the $n \times n$ matrix $[f_i(x_j)]$ is nonsingular.
12. Formulate and solve the isoperimetric problem of Example 1 of Section 7.7 by using polar coordinates.
13. Solve the candidates allocation problem described in Chapter 1.
14. Let $X = L_2[0, 1]$ and define $f(x) = \int_0^1 \{\frac{1}{2}x^2(t) + |x(t)|\}\, dt$ on X. Find the conjugate functional of f.
15. Exhibit a convex set C having the property that the convex conjugate functional of 0 over C is not equal to the negative of the concave conjugate functional of 0 over C.
16. Let M be a nonempty closed convex set in $R \times X$, and assume that there is at least one nonvertical hyperplane containing M in one of its half-spaces. Show that for any $(r_0, x_0) \notin M$ there is a nonvertical hyper-

plane separating (r_0, x_0) and M. Hint: Consider convex combinations of hyperplanes.

17. Let f be a convex functional on a convex set C in a normed space and let $[f^*, C^*] = [f, C]^*$. For $x \in C$, $x^* \in C^*$, deduce Young's inequality

$$\langle x, x^* \rangle \leq f(x) + f^*(x^*).$$

Apply this result to norms in L_p spaces.

18. Derive the minimum norm duality theorem (Theorem 1, Section 5.8) directly from the Fenchel duality theorem.

19. Suppose a fixed quantity x_0 of resource is to be allocated over a given time at n equally spaced time instants. Thus $x_1 \leq x_0$ is allocated first. The remaining resource $x_0 - x_1$ grows by a factor a so that at the second instant $x_2 \leq a(x_0 - x_1)$ may be allocated. In general, the uncommitted resource grows by the factor a between each step. Show that a sequence of allocations $\{x_i\}_{i=1}^n$, $x_i \geq 0$ is feasible if and only if

$$a^{n-1}x_1 + a^{n-2}x_2 + \cdots + ax_{n-1} + x_n \leq a^{n-1}x_0.$$

Hence, show how to generalize the result of Example 1, Section 7.12, to multistage problems.

20. The owner of a small food stand at the beach is about to order his weekend supply of food. Mainly, he sells ice cream and hot dogs and wishes to optimize his allocation of money to these two items. He knows from past experience that the demands for these items depend on the weather in the following way:

	Hot Day	Cool Day
Ice cream	1000	200
Hot dogs	400	200

He believes that a hot or a cool day is equally probable. Anything he doesn't sell he may return for full credit. His profit on ice cream is 10 cents and on hot dogs it is 30 cents. He, of course, wishes to maximize his expected profit while remaining within his budget of \$100. Ice cream costs him 10 cents and hot dogs 20 cents.

(a) Formulate his problem and reduce it to the form

$$\begin{aligned} &\text{maximize } f_1(x_1) + f_2(x_2) \\ &\text{subject to } x_1 + x_2 \leq x_0,\ x_1 \geq 0,\ x_2 \geq 0. \end{aligned}$$

(b) Solve the problem using conjugate functionals.

21. Consider a control system governed by the n-dimensional set of differential equations

$$\dot{x}(t) = Ax(t) + bu(t)$$

which has solution

$$x(T) = \Phi(T)x(0) + \int_0^T \Phi(T - t)bu(t)\, dt,$$

where $\Phi(t)$ is a fundamental matrix of solutions. Using conjugate function theory and duality, find the control u minimizing

$$J = \tfrac{1}{2}\|x(T)\|^2 + \tfrac{1}{2}\int_0^T u^2(t)\, dt.$$

Hint: Assume first that $x(T)$ is known and reduce the problem to a finite-dimensional one. Next optimize $x(T)$. Alternatively, formulate the problem in $E^n \times L_2[0, T]$.

REFERENCES

§7.1–4. For general background on differentials, see Graves [62], [63], Hildebrandt and Graves [72] and Hille and Phillips [73, Chapters 3 and 26]. A fairly comprehensive account of several extensions of ordinary calculus is contained in Luisternik and Sobolev [101]. The inequality extension of the mean value theorem is discussed by Kantorovich and Akilov [79]; also see Antosiewicz and Reinboldt [9].

§7.5–6. The classic treatise on the calculus of variations is Bliss [22]. Also see Fox [55]. For a somewhat more modern approach to the subject, see Gelfand and Fomin [58] or Akhiezer [2].

§7.7. The implicit function theorem used here has generalizations to abstract spaces. See Hildebrandt and Graves [72] and also Section 9.2 of this book. There are a number of abstract versions of the Lagrange multiplier theorem, e.g., Goldstine [60], [61] and Blum [23].

§7.9–12. Conjugate functionals on finite-dimensional space were introduced by Fenchel [52]. For an excellent presentation of this topic, consult Karlin [82]. Some extensions and related discussions are to be found in Rockafellar [126], Rådström [121], and Whinston [151]. Extensions to more general linear topological spaces were made by Brondsted [25], Dieter [41], Rockafellar [127], and Moreau [107]. See Kretschmer [90] for an example of a case where the values of the primal and dual problems are unequal. An equivalent approach, the maximum transform, was developed by Bellman and Karush [19]. They treat the allocation problem. The horse-racing problem is due to Rufus Isaacs and was analyzed in detail by Karlin using the Neyman-Pearson lemma.

§7.13. See Edwards [46], Karlin [81], [82], [83], Dresher [44], and McKinsey [104]. For more general min-max theorems, see Fan [51] and Sion [141].

§7.14. For further developments along the lines of Problem 8, see Varaiya [149]. For a solution to Problem 16, see Brondsted [25].

8

GLOBAL THEORY OF CONSTRAINED OPTIMIZATION

8.1 Introduction

The general optimization problem treated in this book is to locate from within a given subset of a vector space that particular vector which minimizes a given functional. In some problems the subset of admissible vectors competing for the optimum is defined explicitly, as in the case of a given subspace in minimum norm problems; in other cases the subset of admissible vectors is defined implicitly by a set of constraint relations. In previous chapters we considered examples of both types, but generally we reduced a problem with implicit constraints to one with explicit constraints by finding the set of solutions to the constraint relations. In this chapter and the next we make a more complete study of problems with implicit constraints defined by nonlinear equality or inequality relations.

Deservedly dominating our attention, of course, are Lagrange multipliers which somehow almost always unscramble a difficult constrained problem. Although we encountered Lagrange multipliers at several points in previous chapters, they were treated rather naively as a convenient set of numbers or simply as the result of certain duality calculations. In a more sophisticated approach, we do not speak of individual Lagrange multipliers but instead of an entire Lagrange multiplier vector in an appropriate dual space. For example, the problem

$$\text{(1)} \qquad \begin{cases} \text{minimize } f(x) \\ \text{subject to } H(x) = \theta, \end{cases}$$

where H is a mapping from a normed space X into a normed space Z, has Lagrangian

$$L(x, z^*) = f(x) + \langle H(x), z^* \rangle$$

and the Lagrange multiplier is some specific $z^* \in Z^*$. (We also write the Lagrangian in the functional notation

$$L(x, z^*) = f(x) + z^* H(x)$$

since this is similar to the convention for the finite-dimensional theory.)

There are several important geometric interpretations of constrained problems and their corresponding Lagrange multipliers. For example, problem (1) can be viewed in the space X by studying the contours of f, or in $R \times X$ by studying the graph of f. Alternatively, however, the problem can be viewed in the constraint space Z or in $R \times Z$, and these representations are in many respects more illuminating than those in the primal space X because the Lagrange multiplier is an element of Z^* and appears directly as a hyperplane in Z. Lagrange multipliers cannot be thoroughly understood without understanding each of these geometric interpretations.

Because of the interpretation of a Lagrange multiplier as a hyperplane, it is natural to suspect that the theory is simplest and most elegant for problems involving convex functionals. Indeed this is so. In this chapter we therefore consider the global or convex theory based on the geometric interpretation in the constraint space where the Lagrange multiplier appears as a separating hyperplane. In the next chapter we consider the local theory and the geometric interpretation in the primal space X.

8.2 Positive Cones and Convex Mappings

By introducing a cone defining the positive vectors in a given space, it is possible to consider inequality problems in abstract vector spaces.

Definition. Let P be a convex cone in a vector space X. For $x, y \in X$, we write $x \geq y$ (with respect to P) if $x - y \in P$. The cone P defining this relation is called the *positive cone* in X. The cone $N = -P$ is called the *negative cone* in X and we write $y \leq x$ for $y - x \in N$.

For example, in E^n the convex cone

$$P = \{x \in E^n : x = (\xi_1, \xi_2, \ldots, \xi_n);\ \xi_i \geq 0 \text{ for all } i\} \tag{1}$$

defines the ordinary positive orthant. In a space of functions defined on an interval of the real line, say $[t_1, t_2]$, it is natural to define the positive cone as consisting of all functions in the space that are nonnegative everywhere on the interval $[t_1, t_2]$.

We can easily verify that $x \geq y$, $y \geq z$ implies $x \geq z$ and, since $\theta \in P$, $x \geq x$ for all $x \in X$.

In a normed space it is sometimes important to define the positive cone by a closed convex cone. For example, in E^n the cone defined by (1) is closed, but if one or more of the inequalities in the description of the set were changed to strict inequality (and the point θ adjoined), the resulting cone would not be closed.

In the case of a normed space, we write $x > \theta$ if x is an interior point of the positive cone P. For many applications it is essential that P possess an

interior point so that the separating hyperplane theorem can be applied. Nevertheless, this is not possible in many common normed spaces. For instance, if $X = L_1[t_1, t_2]$ and P is taken as the subset of nonnegative functions on the interval $[t_1, t_2]$, we can easily show that P contains no interior point. On the other hand, in $C[t_1, t_2]$ the cone of nonnegative functions does possess interior points; for this reason the space $C[t_1, t_2]$ is of particular interest for problems involving inequalities.

Given a normed space X together with a positive convex cone $P \subset X$, it is natural to define a corresponding positive convex cone $P^{\oplus}$ in the dual space X^* by

$$P^{\oplus} = \{x^* \in X^* : \langle x, x^* \rangle \geq 0 \quad \text{for all } x \in P\}. \tag{2}$$

Example 1. If P is taken as given by (1), then $P^{\oplus}$ consists of all linear functionals represented by elements of E^n with nonnegative components.

Example 2. If in the space $C[t_1, t_2]$ P is taken as the set of all nonnegative continuous functions on $[t_1, t_2]$, then $P^{\oplus}$ consists of all linear functionals on $C[t_1, t_2]$ represented by functions v of bounded variation and nondecreasing on $[t_1, t_2]$.

We can easily show that $P^{\oplus}$ is closed even if P is not. If P is closed, P and $P^{\oplus}$ are related through the following result.

Proposition 1. *Let the positive cone P in the normed space X be closed. If $x \in X$ satisfies*

$$\langle x, x^* \rangle \geq 0 \quad \textit{for all } x^* \geq \theta,$$

then $x \geq \theta$.

Proof. Assume $x \notin P$. Then by the separating hyperplane theorem there is an $x^* \in X^*$ such that $\langle x, x^* \rangle < \langle p, x^* \rangle$ for all $p \in P$. Since P is a cone, $\langle p, x^* \rangle$ can never be negative because then $\langle x, x^* \rangle > \langle \alpha p, x^* \rangle$ for some $\alpha > 0$. Thus $x^* \in P^{\oplus}$. Also, since $\inf_{p \in P} \langle p, x^* \rangle = 0$, we have $\langle x, x^* \rangle < 0$. ∎

Since we have generalized the notion of inequality between vectors, it is possible to introduce a general definition of convexity for mappings.

Definition. Let X be a vector space and let Z be a vector space having a cone P specified as the positive cone. A mapping $G: X \to Z$ is said to be *convex* if the domain Ω of G is a convex set and if $G(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha G(x_1) + (1 - \alpha)G(x_2)$ for all $x_1, x_2 \in \Omega$ and all α, $0 < \alpha < 1$.

We note that convexity is not an intrinsic property of a mapping but is dependent on the specified positive cone in the range space.

The following elementary property of convex mappings is obviously of particular interest in constrained optimization problems.

Proposition 2. *Let G be a convex mapping as in the last definition. Then for every $z \in Z$ the set $\{x : x \in \Omega, G(x) \leq z\}$ is convex.*

8.3 Lagrange Multipliers

The basic problem considered in the next few sections is:

$$\text{(1)} \qquad \begin{cases} \text{minimize } f(x) \\ \text{subject to } x \in \Omega,\ G(x) \leq \theta, \end{cases}$$

where Ω is a convex subset of a vector space X, f is a real-valued convex functional on Ω, and G is a convex mapping from Ω into a normed space Z having positive cone P. Problem (1) above is referred to as the general convex programming problem.

We analyze problem (1) and develop the Lagrange multiplier theorem essentially by imbedding it in the family of problems

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in \Omega,\ G(x) \leq z \end{aligned}$$

where z is an arbitrary vector in Z. The solution to these problems depends on z and it is an examination of this dependency that guides our analysis.

In view of the above remark, we define the set $\Gamma \subset Z$ as

$$\Gamma = \{z: \text{There is an } x \in \Omega \text{ with } G(x) \leq z\}.$$

The set Γ is convex since $z_1, z_2 \in \Gamma$ implies the existence of $x_1, x_2 \in \Omega$ with $G(x_1) \leq z_1$, $G(x_2) \leq z_2$; hence, for any α, $0 < \alpha < 1$, $G(\alpha x_1 + (1-\alpha)x_2) \leq \alpha z_1 + (1-\alpha)z_2$ which implies $\alpha z_1 + (1-\alpha)z_2 \in \Gamma$.

On the set Γ, we define the *primal functional* ω (which may not be finite) as

$$\omega(z) = \inf \{f(x) : x \in \Omega,\ G(x) \leq z\}.$$

The original problem (1) can be regarded as determining the single value $\omega(\theta)$. The entire theory of this chapter is based on a study of ω.

Proposition 1. *The functional ω is convex.*

Proof.

$$\begin{aligned} \omega(\alpha z_1 + (1-\alpha)z_2) &= \inf \{f(x) : x \in \Omega,\ G(x) \leq \alpha z_1 + (1-\alpha)z_2\} \\ &\leq \inf \{f(x) : x = \alpha x_1 + (1-\alpha)x_2,\ x_1 \in \Omega,\ x_2 \in \Omega, \\ &\qquad G(x_1) \leq z_1,\ G(x_2) \leq z_2\} \\ &\leq \alpha \inf \{f(x_1) : x_1 \in \Omega,\ G(x_1) \leq z_1\} \\ &\qquad + (1-\alpha) \inf \{f(x_2) : x_2 \in \Omega,\ G(x_2) \leq z_2\} \\ &\leq \alpha\omega(z_1) + (1-\alpha)\omega(z_2). \ \blacksquare \end{aligned}$$

Proposition 2. *The functional ω is decreasing; i.e., if $z_1 \geq z_2$, then $\omega(z_1) \leq \omega(z_2)$.*

Proof. This is immediate. ∎

A typical ω for Z one dimensional is shown in Figure 8.1.

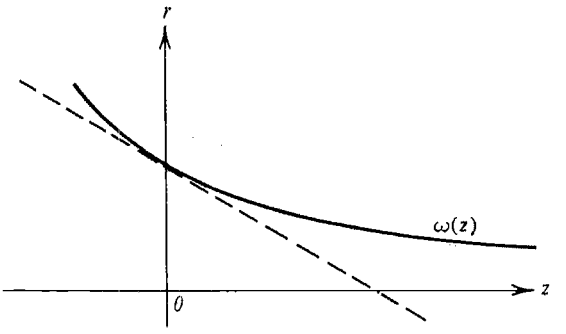

Figure 8.1 The primal functional

Conceptually the Lagrange multiplier theorem follows from the simple observation that since ω is convex, there is a hyperplane tangent to ω at $z = \theta$ and lying below ω throughout its region of definition. If one were to tilt his head so that the tangent hyperplane became the new horizontal, it would appear that ω was minimized at θ or, said another way, by adding an appropriate linear functional $\langle z, z_0^* \rangle$ to $\omega(z)$, the resulting combination $\omega(z) + \langle z, z_0^* \rangle$ is minimized at $z = \theta$. The functional z_0^* is the Lagrange multiplier for the problem; the tangent hyperplane illustrated in Figure 8.1 corresponds to the element $(1, z_0^*) \in R \times Z^*$.

The discussion above is made precise by the following theorem.

Theorem 1. *Let X be a linear vector space, Z a normed space, Ω a convex subset of X, and P the positive cone in Z. Assume that P contains an interior point.*

Let f be a real-valued convex functional on Ω and G a convex mapping from Ω into Z. Assume the existence of a point $x_1 \in \Omega$ for which $G(x_1) < \theta$ (i.e., $G(x_1)$ is an interior point of $N = -P$).

Let

$$\mu_0 = \inf f(x) \qquad \text{subject to } x \in \Omega,\ G(x) \leq \theta \tag{2}$$

and assume μ_0 is finite. Then there is an element $z_0^* \geq \theta$ in Z^* such that

$$\mu_0 = \inf_{x \in \Omega} \{f(x) + \langle G(x), z_0^* \rangle\}. \tag{3}$$

Furthermore, if the infimum is achieved in (2) *by an* $x_0 \in \Omega$, $G(x_0) \leq \theta$, *it is achieved by* x_0 *in* (3) *and*

$$\langle G(x_0), z_0^* \rangle = 0 \tag{4}$$

Proof. In the space $W = R \times Z$, define the sets

$$A = \{(r, z) : r \geq f(x), z \geq G(x) \text{ for some } x \in \Omega\}$$

$$B = \{(r, z) : r \leq \mu_0, z \leq \theta\}.$$

Since f and G are convex, both A and B are convex sets. (It should be noted that the set A is the convex region above the graph of the primal functional ω.) The definition of μ_0 implies that A contains no interior points of B. Also, since N contains an interior point, the set B contains an interior point. Thus, according to the separating hyperplane theorem, there is a nonzero element $w_0^* = (r_0, z_0^*) \in W^*$ such that

$$r_0 r_1 + \langle z_1, z_0^* \rangle \geq r_0 r_2 + \langle z_2, z_0^* \rangle$$

for $(r_1, z_1) \in A$, $(r_2, z_2) \in B$.

From the nature of B it follows immediately that $w_0^* \geq \theta$ or, equivalently, that $r_0 \geq 0$, $z_0^* \geq \theta$. We now show that $r_0 > 0$. The point (μ_0, θ) is in B; hence

$$r_0 r + \langle z, z_0^* \rangle \geq r_0 \mu_0$$

for all $(r, z) \in A$. If r_0 were zero, it would follow in particular that $\langle G(x_1), z_0^* \rangle \geq 0$ and that $z_0^* \neq \theta$. However, since $G(x_1)$ is an interior point of N and $z_0^* \geq \theta$, it follows that $\langle G(x_1), z_0^* \rangle < 0$ (the reader should verify this), which is a contradiction. Therefore, $r_0 > 0$ and, without loss of generality, we may assume $r_0 = 1$.

Since the point (μ_0, θ) is arbitrarily close to A and B, we have (with $r_0 = 1$)

$$\mu_0 = \inf_{(r, z) \in A} [r + \langle z, z_0^* \rangle] \leq \inf_{x \in \Omega} [f(x) + \langle G(x), z_0^* \rangle]$$

$$\leq \inf_{\substack{x \in \Omega \\ G(x) \leq \theta}} f(x) = \mu_0.$$

Hence the first part of the theorem is proved.

If there exists an x_0 such that $G(x_0) \leq \theta$, $\mu_0 = f(x_0)$, then

$$\mu_0 \leq f(x_0) + \langle G(x_0), z_0^* \rangle \leq f(x_0) = \mu_0$$

and hence $\langle G(x_0), z_0^* \rangle = 0$. ∎

The proof of Theorem 1 depends partially on the convexity of set A. This set, being the region above the graph of the primal functional ω, is

convex if and only if ω is convex. This in turn is implied by, but weaker than, the assumption of convexity of f and G.

Aside from convexity there are two important assumptions in Theorem 1 that deserve comment: the assumption that the positive cone P contains an interior point and the assumption that $G(x_1) < \theta$ for some $x_1 \in \Omega$. These assumptions are introduced to guarantee the existence of a nonvertical separating hyperplane. The requirement that the positive cone possess an interior point is fairly severe and apparently cannot be completely omitted. Indeed, in many applications this requirement is the determining factor in the choice of space in which a problem is formulated and, of course, $C[a, b]$ is a natural candidate for problems involving functions on the interval $[a, b]$. The condition $G(x_1) < \theta$ is called a regularity condition and is typical of the assumptions that must be made in Lagrange multiplier theorems. It guarantees that the separating hyperplane is nonvertical.

We have considered only convex constraints of the form $G(x) \leq \theta$. An equality constraint $H(x) = \theta$ with $H(x) = Ax - b$, where A is linear, is equivalent to the two convex inequalities $H(x) \leq \theta$ and $-H(x) \leq \theta$ and can thus be cast into the form $G(x) \leq \theta$. One expects then that, as a trivial corollary to Theorem 1, linear equality constraints can be included together with their resulting Lagrange multipliers. There is a slight difficulty, however, since there never exists an x_1 which simultaneously renders $H(x_1) < \theta$ and $-H(x_1) < \theta$. A composite theorem for inequality constraints and a finite number of equality constraints is given in Problem 7.

Theorem 1 is a geometric version of the Lagrange multiplier theorem for convex problems. An equivalent algebraic formulation of the results is given by the following saddle-point statement.

Corollary 1. *Let everything be as in Theorem 1 and assume that x_0 achieves the constrained minimum. Then there is a $z_0^* \geq \theta$ such that the Lagrangian*

$$L(x, z^*) = f(x) + \langle G(x), z^* \rangle$$

has a saddle point at x_0, z_0^*; i.e.,

$$L(x_0, z^*) \leq L(x_0, z_0^*) \leq L(x, z_0^*)$$

for all $x \in \Omega$, $z^* \geq \theta$.

Proof. Let z_0^* be defined as in Theorem 1. From (3) we have immediately $L(x_0, z_0^*) \leq L(x, z_0^*)$. By equation (4) we have

$$L(x_0, z^*) - L(x_0, z_0^*) = \langle G(x_0), z^* \rangle - \langle G(x_0), z_0^* \rangle = \langle G(x_0), z^* \rangle \leq 0. \quad \blacksquare$$

8.4 Sufficiency

The conditions of convexity and existence of interior points cannot be omitted if we are to guarantee the existence of a separating hyperplane in the

space $R \times Z$. If, however, the appropriate hyperplane does exist, despite the absence of these conditions, the Lagrange technique for location of the optimum still applies. The situation is illustrated in Figure 8.2, where again

$$\omega(z) = \inf \{f(x) : x \in \Omega, G(x) \leq z\}$$

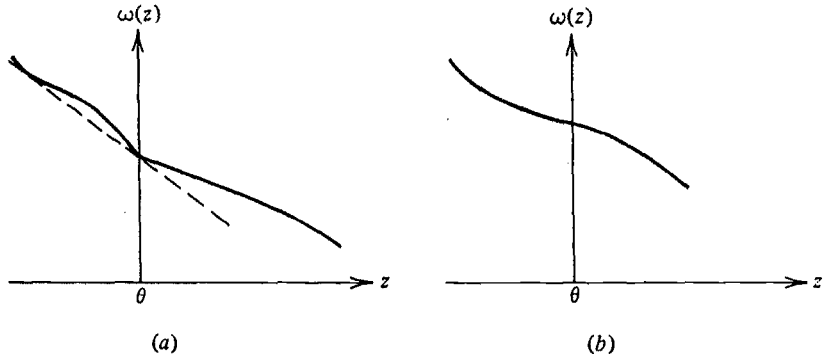

Figure 8.2 Nonconvexity

is plotted, but the convexity of ω is not assumed. If, as in Figure 8.2*a*, an appropriate hyperplane exists, it is fairly clear that $f(x) + \langle G(x), z_0^* \rangle$ attains a minimum at x_0. In Figure 8.2*b*, no supporting hyperplane exists at $z = \theta$ and the Lagrange statement cannot be made.

These observations lead to the following sufficiency theorems.

Theorem 1. *Let f be a real-valued functional defined on a subset Ω of a linear space X. Let G be a mapping from Ω into the normed space Z having nonempty positive cone P.*

Suppose there exists an element $z_0^* \in Z^*$, $z_0^* \geq \theta$, and an element $x_0 \in \Omega$ such that

$$f(x_0) + \langle G(x_0), z_0^* \rangle \leq f(x) + \langle G(x), z_0^* \rangle$$

for all $x \in \Omega$. Then x_0 solves:

$$\begin{aligned} &\textit{minimize } f(x) \\ &\textit{subject to } G(x) \leq G(x_0), \qquad x \in \Omega. \end{aligned}$$

Proof. Suppose there is an $x_1 \in \Omega$ with $f(x_1) < f(x_0)$, $G(x_1) \leq G(x_0)$. Then, since $z_0^* \geq \theta$, it follows that

$$\langle G(x_1), z_0^* \rangle \leq \langle G(x_0), z_0^* \rangle$$

and hence that

$$f(x_1) + \langle G(x_1), z_0^* \rangle < f(x_0) + \langle G(x_0), z_0^* \rangle$$

which contradicts the hypothesis of the theorem. ∎

Theorem 2. *Let X, Z, Ω, P, f, G be as above and assume that P is closed. Suppose there exists a $z_0^* \in Z^*$, $z_0^* \geq \theta$, and an $x_0 \in \Omega$ such that the Lagrangian $L(x, z^*) = f(x) + \langle G(x), z^* \rangle$ possess a saddle point at x_0, z_0^*; i.e.,*

$$L(x_0, z^*) \leq L(x_0, z_0^*) \leq L(x, z_0^*),$$

for all $x \in \Omega$, $z^* \geq \theta$. Then x_0 solves:

$$\begin{aligned} &\textit{minimize } f(x) \\ &\textit{subject to } G(x) \leq \theta, \qquad x \in \Omega. \end{aligned}$$

Proof. The saddle-point condition with respect to z^* gives

$$\langle G(x_0), z^* \rangle \leq \langle G(x_0), z_0^* \rangle$$

for all $z^* \geq \theta$. Hence, in particular, for all $z_1^* \geq \theta$

$$\langle G(x_0), z_1^* + z_0^* \rangle \leq \langle G(x_0), z_0^* \rangle$$

or

$$\langle G(x_0), z_1^* \rangle \leq 0.$$

We conclude by Proposition 1, Section 8.2, that $G(x_0) \leq \theta$. The saddle-point condition therefore implies that $\langle G(x_0), z_0^* \rangle = 0$.

Assume now that $x_1 \in \Omega$ and that $G(x_1) \leq \theta$. Then, according to the saddle-point condition with respect to x,

$$f(x_0) = f(x_0) + \langle G(x_0), z_0^* \rangle \leq f(x_1) + \langle G(x_1), z_0^* \rangle \leq f(x_1).$$

Thus x_0 minimizes $f(x)$ subject to $x \in \Omega$, $G(x) \leq \theta$. ∎

The saddle-point condition offers a convenient compact description of the essential elements of the Lagrangian results for convex programming. If f and G are convex, the positive cone $P \subset Z$ is closed and has nonempty interior, and the regularity condition is satisfied, then the saddle-point condition is necessary and sufficient for optimality of x_0.

8.5 Sensitivity

The Lagrange theorems of the preceding sections do not exploit all of the geometric properties evident from the representation of the problem in $R \times Z$. Two other properties, sensitivity and duality, important for both theory and application, are obtainable by visualization of the functional ω.

For any z_0 the hyperplane determined by the Lagrange multiplier for the problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } G(x) \leq z_0, \qquad x \in \Omega \end{aligned}$$

is a support hyperplane at $\omega(z_0)$, and this hyperplane serves as a lower bound for ω as illustrated in Figure 8.3. This observation produces the following sensitivity theorem.

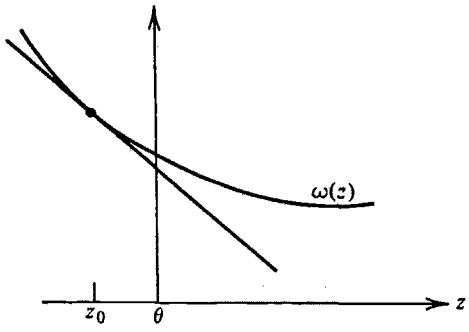

Figure 8.3 Sensitivity

Theorem 1. *Let f and G be convex and suppose x_0, x_1 are solutions to the problems*

$$\begin{aligned} &\textit{minimize } f(x) \\ &\textit{subject to } x \in \Omega \quad \textit{and} \quad G(x) \leq z_0, \qquad G(x) \leq z_1, \end{aligned}$$

respectively. Suppose z_0^*, z_1^* are Lagrange multipliers corresponding to these problems. Then

$$\langle z_1 - z_0, z_1^* \rangle \leq f(x_0) - f(x_1) \leq \langle z_1 - z_0, z_0^* \rangle.$$

Proof. The Lagrange multiplier z_0^* makes

$$f(x_0) + \langle G(x_0) - z_0, z_0^* \rangle \leq f(x) + \langle G(x) - z_0, z_0^* \rangle,$$

for all $x \in \Omega$. In particular, setting $x = x_1$ and taking account of $\langle G(x_0) - z_0, z_0^* \rangle = 0$ yields

$$f(x_0) - f(x_1) \leq \langle G(x_1) - z_0, z_0^* \rangle \leq \langle z_1 - z_0, z_0^* \rangle.$$

A similar argument applied to x_1, z_1^* produces the other inequality. ∎

A statement equivalent to that of Theorem 1 is that

$$\omega(z) - \omega(z_0) \geq \langle z_0 - z, z_0^* \rangle.$$

Hence, if the functional ω is Fréchet differentiable at z_0, it follows (see Problem 9) that

$$\omega'(z_0) = -z_0^*.$$

Therefore, the Lagrange multiplier z_0^* is the negative of the first-order sensitivity of the optimal objective with respect to the constraint term z_0.

8.6 Duality

Consider again the basic convex programming problem:

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } G(x) \leq \theta, \qquad x \in \Omega \end{aligned}$$

where f, G and Ω are convex. The general duality principle for this problem is based on the simple geometric properties of the problem viewed in $R \times Z$. As in Section 8.3, we define the primal functional on the set Γ

$$\omega(z) = \inf \{f(x) : G(x) \leq z,\ x \in \Omega\} \tag{1}$$

and let $\mu_0 = \omega(\theta)$. The duality principle is based on the observation that (as illustrated in Figure 8.4) μ_0 is equal to the maximum intercept with the

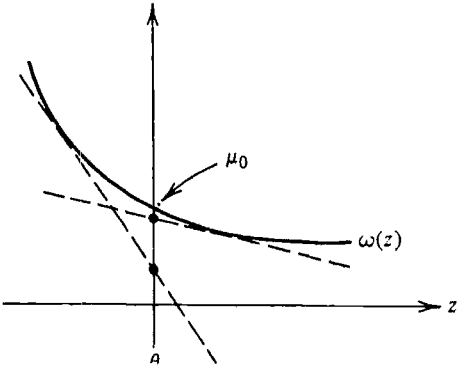

Figure 8.4 Duality

vertical axis of all closed hyperplanes that lie below ω. The maximum intercept is, of course, attained by the hyperplane determined by the Lagrange multiplier of the problem.

To express the above duality principle analytically, we introduce the *dual functional* φ corresponding to (1) to be defined on the positive cone in Z^* as

$$\varphi(z^*) = \inf_{x \in \Omega} [f(x) + \langle G(x), z^* \rangle]. \tag{2}$$

In general, φ is not finite throughout the positive cone in Z^* but the region where it is finite is convex.

Proposition 1. *The dual functional is concave and can be expressed as*

$$\varphi(z^*) = \inf_{z \in \Gamma} [\omega(z) + \langle z, z^* \rangle]. \tag{3}$$

Proof. The concavity of φ is easy to show and is left to the reader. For any $z^* \geq \theta$ and any $z \in \Gamma$, we have

$$\varphi(z^*) = \inf_{x \in \Omega} [f(x) + \langle G(x), z^* \rangle] \leq \inf \{f(x) + \langle z, z^* \rangle : G(x) \leq z, x \in \Omega\}$$
$$= \omega(z) + \langle z, z^* \rangle.$$

On the other hand, for any $x_1 \in \Omega$ we have, with $z_1 = G(x_1)$,

$$f(x_1) + \langle G(x_1), z^* \rangle \geq \inf \{f(x) + \langle z_1, z^* \rangle : G(x) \leq z_1, x \in \Omega\}$$
$$= \omega(z_1) + \langle z_1, z^* \rangle$$

and hence

$$\varphi(z^*) \geq \inf_{z \in \Gamma} [\omega(z) + \langle z, z^* \rangle].$$

Therefore, equality must hold in (3). ∎

The element $(1, z^*) \in R \times Z^*$ determines a family of hyperplanes in $R \times Z$, each hyperplane consisting of the points (r, z) which satisfy $r + \langle z, z^* \rangle = k$ where k is a constant. Equation (3) says that for $k = \varphi(z^*)$ this hyperplane supports the set $[\omega, \Gamma]$, the region above the graph of ω. Furthermore, at $z = \theta$ we have $r = \varphi(z^*)$; hence, $\varphi(z^*)$ is equal to the intercept of this hyperplane with the vertical axis. It is clear then that the dual functional is precisely what is needed to express the duality principle.

Referring to Section 7.10 we see that φ is essentially the conjugate functional of ω. Unfortunately, as is often the case with theories that are developed independently and later found to be intimately linked, there is a discrepancy in sign convention between the dual functional and the conjugate functional but the essential concepts are identical.

In view of the above discussion the following result establishing the equivalence of two extremization problems—the minimization of a convex functional and the maximization of a concave functional—should be self-evident.

Theorem 1. (*Lagrange Duality*) *Let f be a real-valued convex functional defined on a convex subset Ω of a vector space X, and let G be a convex mapping of X into a normed space Z. Suppose there exists an x_1 such that $G(x_1) < \theta$ and that $\mu_0 = \inf \{f(x) : G(x) \leq \theta, x \in \Omega\}$ is finite. Then*

$$\inf_{\substack{G(x) \leq \theta \\ x \in \Omega}} f(x) = \max_{z^* \geq \theta} \varphi(z^*) \tag{4}$$

and the maximum on the right is achieved by some $z_0^* \geq \theta$.

If the infimum on the left is achieved by some $x_0 \in \Omega$, then

$$\langle G(x_0), z_0^* \rangle = 0$$

and x_0 minimizes $f(x) + \langle G(x), z_0^* \rangle$, $x \in \Omega$.

Proof. For any $z^* \geq \theta$ we have

$$\inf_{x \in \Omega} [f(x) + \langle G(x), z^* \rangle] \leq \inf_{\substack{x \in \Omega \\ G(x) \leq \theta}} [f(x) + \langle G(x), z^* \rangle] \leq \inf_{\substack{x \in \Omega \\ G(x) \leq \theta}} f(x) = \mu_0 .$$

Therefore, the right-hand side of the equation in the theorem statement is less than or equal to μ_0. However, Theorem 1, Section 8.3, establishes the existence of an element z_0^* which gives equality. The remainder of the theorem statement is given in Theorem 1, Section 8.3. ∎

Since ω is decreasing, $\omega(\theta) \leq \omega(z)$ for all $z \leq \theta$; hence, an alternative symmetric formulation of (4) which emphasizes the duality (but which is one step removed from the original problem), is

$$\min_{z \leq \theta} \omega(z) = \max_{z^* \geq \theta} \varphi(z^*). \tag{5}$$

As a final note we observe that even in nonconvex programming problems the dual functional can be formed according to (2). From the geometric interpretation of φ in terms of hyperplanes supporting $[\omega, \Gamma]$, it is clear that φ formed according to (2) will be equal to that which would be obtained by considering hyperplanes supporting the convex hull of $[\omega, \Gamma]$. Also from this interpretation it follows (provided $\varphi(z^*)$ is finite for some $z^* \geq \theta$) that for any programming problem

$$\max_{z^* \geq \theta} \varphi(z^*) \leq \min_{z \leq \theta} \omega(z) \tag{6}$$

and hence the dual functional always serves as a lower bound to the value of the primal problem.

Example 1. (Quadratic Programming) As a simple application of the duality theorem, we calculate the dual of the quadratic program:

$$\begin{aligned} &\text{minimize } \tfrac{1}{2} x'Qx - b'x \\ &\text{subject to } Ax \leq c \end{aligned}$$

where x is an n vector to be determined, b is an n vector, Q is an $n \times n$ positive-definite symmetric matrix, A is an $m \times n$ matrix, and c is an m vector.

Assuming feasibility (i.e., assuming there is an x satisfying $Ax < c$), the problem is equivalent to

$$\max_{\lambda \geq \theta} \min_{x} \{\tfrac{1}{2} x'Qx - b'x + \lambda'[Ax - c]\}. \tag{7}$$

The minimization over x is unconstrained and is attained by

$$x = Q^{-1}(b - A'\lambda).$$

Substituting this in Problem (7), the problem becomes

$$\max_{\lambda \geq \theta} \{-\tfrac{1}{2}\lambda' P\lambda - \lambda' d - \tfrac{1}{2}b'Q^{-1}b\},$$

where

$$P = AQ^{-1}A', \qquad d = c - AQ^{-1}b.$$

Thus the dual is also a quadratic programming problem. Note that the dual problem may be much easier to solve than the primal problem since the constraint set is much simpler and the dimension is smaller if $m < n$.

Example 2. (Quadratic Programming) Consider now the quadratic programming problem:

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}x'Qx - b'x \\ &\text{subject to } Ax \leq c, \end{aligned}$$

where Q is assumed to be only positive semidefinite rather than positive definite. The dual is again

$$\max_{\lambda \geq \theta} \min_{x} \{\tfrac{1}{2}x'Qx - b'x + \lambda'[Ax - c]\}, \tag{8}$$

but now the minimum over x may not be finite for every $\lambda \geq \theta$. In fact, the minimization over x is finite if and only if there is an x satisfying

$$Qx - b + A'\lambda = \theta$$

and all x's satisfying this equation yield the minimum. With this equation substituted in (8), we obtain the dual in the form

$$\begin{aligned} &\text{maximize } -\lambda'c - \tfrac{1}{2}x'Qx \\ &\text{subject to } Qx - b + A'\lambda = \theta, \qquad \lambda \geq \theta, \end{aligned}$$

where the maximum is taken with respect to both λ and x.

8.7 Applications

In this section we use the theory developed in this chapter to solve several allocation and control problems involving inequality constraints.

Although the theory of convex programming developed earlier does not require even continuity of the convex functionals involved, in most problems the functionals are not only continuous but Fréchet differentiable. In such problems it is convenient to express necessary or sufficient conditions

in differential form. For this reason we make frequent use of the following lemma which generalizes the observation that if a convex function f on $[0, \infty)$ achieves a minimum at an interior point x_0, then $f'(x_0) = 0$, while if it achieves a minimum at $x_0 = 0$, then $f'(x_0) \geq 0$; in either case $x_0 f'(x_0) = 0$.

Lemma 1. *Let f be a Fréchet differentiable convex functional on a real normed space X. Let P be a convex cone in X. A necessary and sufficient condition that $x_0 \in P$ minimize f over P is that*

$$\delta f(x_0; x) \geq 0 \qquad \text{all } x \in P \tag{1}$$

$$\delta f(x_0; x_0) = 0. \tag{2}$$

Proof. Necessity: If x_0 minimizes f, then for any $x \in P$ we must have

$$\frac{d}{d\alpha} f(x_0 + \alpha(x - x_0))\Big|_{\alpha=0} \geq 0.$$

Hence

$$\delta f(x_0; x - x_0) \geq 0. \tag{3}$$

Setting $x = x_0/2$ yields

$$\delta f(x_0; x_0) \leq 0, \tag{4}$$

while setting $x = 2x_0$ yields

$$\delta f(x_0; x_0) \geq 0. \tag{5}$$

Together, equations (3), (4), and (5) imply (1) and (2).

Sufficiency: For x_0, $x \in P$ and $0 < \alpha < 1$ we have

$$f(x_0 + \alpha(x - x_0)) \leq f(x_0) + \alpha[f(x) - f(x_0)]$$

or

$$f(x) - f(x_0) \geq \frac{1}{\alpha}[f(x_0 + \alpha(x - x_0)) - f(x_0)].$$

As $\alpha \to 0+$, the right side of this equation tends toward $\delta f(x_0; x - x_0)$; hence we have

$$f(x) - f(x_0) \geq \delta f(x_0; x - x_0). \tag{6}$$

If (1) and (2) hold, then $\delta f(x_0; x - x_0) \geq 0$ for all $x \in P$ and, hence, from (6)

$$f(x) - f(x_0) \geq 0$$

for all $x \in P$. ∎

Note that the Fréchet differentials can be replaced everywhere by Gateaux differentials in the lemma, provided the Gateaux differentials are linear.

In many of the applications considered in this section, the unknown vector x is constrained to be positive so the constraint set Ω is a cone. In these cases, Lemma 1 is used to express the condition that $f(x) + \langle G(x), z^* \rangle$ is minimized.

Example 1. (Optimal Control) Consider a system governed by the set of differential equations

$$\dot{x}(t) = A(t)x(t) + b(t)u(t), \tag{7}$$

where $x(t)$ is an $n \times 1$ state vector, $A(t)$ is an $n \times n$ matrix, $b(t)$ is an $n \times 1$ distribution matrix, and $u(t)$ is a scalar control.

Given the initial state $x(t_0)$, we seek the control u_0 minimizing

$$J = \tfrac{1}{2} \int_{t_0}^{t_1} u^2(t)\, dt, \tag{8}$$

while satisfying the terminal inequalities

$$x(t_1) \geq c,$$

where c is a fixed $n \times 1$ vector and $t_1 \geq t_0$ is fixed. This problem might represent the selection of a thrust program for a rocket which must exceed certain altitude and velocity limits in a given time or the selection of a production program for a plant with constraints on the total amount produced in a given time.

We can write the solution to equation (7) in the form

$$x(t_1) = \Phi(t_1, t_0)x(t_0) + \int_{t_0}^{t_1} \Phi(t_1, t)b(t)u(t)\, dt, \tag{9}$$

where Φ is the fundamental solution matrix of the corresponding homogeneous equation. We assume $\Phi(t_1, t)$ and $b(t)$ to be continuous. The original problem can now be expressed as: minimize

$$J = \tfrac{1}{2} \int_{t_0}^{t_1} u^2(t)\, dt$$

subject to

$$Ku \geq d, \tag{10}$$

where $d = c - \Phi(t_1, t_0)x(t_0)$ and K is the integral operator defined as

$$Ku = \int_{t_0}^{t_1} \Phi(t_1, t)b(t)u(t)\, dt.$$

This is a convex programming problem defined on, say, $L_2[t_0, t_1]$ with the constraint space being finite dimensional. Using the duality theorem, this infinite-dimensional problem can be reduced to a finite-dimensional one.[1]

Denoting the minimum of (8) under the constraints (10) by μ_0, the duality theorem gives

$$\mu_0 = \max_{\lambda \geq \theta} \min_{u} \{J(u) + \lambda'(d - Ku)\},$$

where λ is an $n \times 1$ vector. More explicitly,

$$\mu_0 = \max_{\lambda \geq \theta} \min_{u} \int_{t_0}^{t_1} [\tfrac{1}{2}u^2(t) - \lambda'\Phi(t_1, t)b(t)u(t)]\, dt + \lambda' d \tag{11}$$

and hence

$$\mu_0 = \max_{\lambda \geq \theta} \lambda' Q\lambda + \lambda' d, \tag{12}$$

where

$$Q = -\tfrac{1}{2} \int_{t_0}^{t_1} \Phi(t_1, t)b(t)b'(t)\Phi'(t_1, t)\, dt.$$

Problem (12) is a simple finite-dimensional maximization problem. Once the solution λ_0 is determined, the optimal control $u_0(t)$ is then given by the function that minimizes the corresponding term in (11). Thus

$$u_0(t) = \lambda_0'\Phi(t_1, t)b(t).$$

The Lagrange multiplier vector λ_0 has the usual interpretation as a sensitivity. In this case it is the gradient of the optimal cost with respect to the target c.

Example 2. (Oil Drilling) A company has located an underground oil deposit of known quantity α. It wishes to determine the long-term oil extraction program that will maximize its total discounted profit.

The problem may be formulated as that of finding the integrable function x on $[0, \infty)$, representing the extraction rate, that maximizes

$$\int_0^\infty F[x(t)]v(t)\, dt$$

subject to

$$\int_0^\infty x(t)\, dt \leq \alpha, \qquad x(t) \geq 0.$$

[1] A problem almost identical to this is solved in Section 7.12 by use of the Fenchel duality theorem.

$F[x]$ represents the profit rate associated with the extraction rate x. By assuming diminishing marginal returns, the function F can be argued to be strictly concave and increasing on $[0, \infty)$ with $F[0] = 0$. We assume also that F has a continuous derivative. The function $v(t)$ represents the discount factor and can be assumed to be continuous, positive, strictly decreasing toward zero, and integrable on $[0, \infty)$.

To apply the differentiability conditions, we first consider the problem with the additional restriction that x be continuous and $x(t) = 0$ for $t \geq T$ where T is some fixed positive time. The constraint space Z then corresponds to the inequality $\int_0^T x\,dt \leq \alpha$ and is thus only one dimensional. The Lagrangian for this problem is

$$L(x, \lambda) = \int_0^T F[x(t)]v(t)\,dt - \lambda\left[\int_0^T x(t)\,dt - \alpha\right] \tag{13}$$

where the minus sign is used because we are maximizing a concave functional. In this case we want $x_0(t) \geq 0$ and λ_0 such that

$$\min_{\lambda \geq 0} \max_{x \geq \theta} L(x, \lambda) = L(x_0, \lambda_0).$$

In view of the concavity of F and Lemma 1, maximization of $L(x, \lambda_0)$ over $x \geq \theta$ is equivalent to

$$\int_0^T \{F_x[x_0(t)]v(t) - \lambda_0\}x(t)\,dt \leq 0 \tag{14}$$

for all $x(t) \geq 0$, and

$$\int_0^T \{F_x[x_0(t)]v(t) - \lambda_0\}x_0(t)\,dt = 0 \tag{15}$$

where F_x is the derivative of F.

It is clear that the solution will have $\lambda_0 > 0$; therefore we seek $x_0(t)$, λ_0, $\mu(t)$ satisfying, for $t \in [0, T]$,

$$x_0(t) \geq 0, \quad \lambda_0 > 0, \quad \mu(t) \geq 0, \quad \mu(t)x_0(t) = 0, \quad \int_0^T x_0(t)\,dt = \alpha, \tag{16}$$

$$F_x[x_0(t)]v(t) - \lambda_0 + \mu(t) = 0. \tag{17}$$

Since F_x is continuous and strictly decreasing, it has an inverse F_x^{-1} which is continuous and strictly decreasing. From (16) and (17) we see immediately that on any interval where $x_0(t) > 0$ we have $x_0(t) = F_x^{-1}\{\lambda_0/v(t)\}$ which is decreasing. Since we seek a continuous x, it follows that the maximizing x_0, if it exists, must have the form

$$x_0(t) = \begin{cases} F_x^{-1}\left\{\dfrac{\lambda_0}{v(t)}\right\} & 0 \leq t \leq t_0 \\ 0 & t_0 \leq t \leq T. \end{cases}$$

In this solution we may have $t_0 = T$ or, if $t_0 < T$, continuity demands that $F_x^{-1}\{\lambda_0/v(t_0)\} = 0$ and hence that $\lambda_0 = F_x[0]v(t_0)$.

Defining

$$J(t_0) = \int_0^{t_0} F_x^{-1}\left\{\frac{F_x[0]v(t_0)}{v(t)}\right\} dt,$$

one can show (we leave it to the reader) that J is continuous and that $\lim_{t_0 \to 0} J(t_0) = 0$ and $\lim_{t_0 \to \infty} J(t_0) = \infty$. Hence, let us choose t_0 such that $J(t_0) = \alpha$. Then for $T > t_0$, the solution

$$x_0(t) = \begin{cases} F_x^{-1}\left\{\dfrac{F_x[0]v(t_0)}{v(t)}\right\} & 0 \le t \le t_0 \\ 0 & t_0 \le t \le T \end{cases}$$

$$\lambda_0 = F_x[0]v(t_0)$$

$$\mu(t) = \begin{cases} 0 & 0 \le t \le t_0 \\ \lambda_0 - F_x[0]v(t) & t_0 \le t \le T \end{cases}$$

satisfies (16) and (17). Since the Lagrangian is concave, these conditions imply that x_0 is optimal for the problem on $[0, T]$.

Given $\varepsilon > 0$, however, any function $y(t)$ for which

$$\int_0^\infty y(t)\, dt = \alpha < \infty$$

can be approximated by a continuous function $x(t)$ vanishing outside a finite interval in such a way that

$$\int_0^\infty \{F[y(t)] - F[x(t)]\}\, v(t)dt < \varepsilon$$

and hence $x_0(t)$ is optimal with respect to this larger class.

The Lagrange multiplier λ_0 in this case is the derivative of the total discounted profit with respect to α.

Example 3. (The Farmer's Allocation Problem) A farmer produces a single crop such as wheat. After harvesting his crop, he may store it or sell and reinvest it by buying additional land and equipment to increase his production rate. The farmer wishes to maximize the total amount stored up to time T.

Letting

$x_1(t)$ = rate of production

$x_2(t)$ = rate of reinvestment

$x_3(t)$ = rate of storage

changing the constraint

$$x_2(t) \leq x_1(t)$$

to

$$x_2(t) \leq x_1(t) + \delta(t - \tau)$$

where δ is the delta function. According to the sensitivity result (which is exact for this problem), the corresponding change in the objective, total storage at $t = T$, is

$$\Delta J = \int_0^T \delta(t - \tau)\, dv(t) = \left.\frac{dv(t)}{dt}\right|_{t=\tau}$$

Thus $dv/dt|_{t=\tau}$ represents the value, in units of final storage, of an opportunity to reinvest one unit of storage at time τ.

Example 4. (Production Planning) This example illustrates that the interior point assumptions in the Lagrange theorem can be important in even simple problems. Consider the following problem:

$$\text{minimize } \tfrac{1}{2}\int_0^1 r^2(t)\, dt$$

$$\text{subject to } \dot{z}(t) = r(t), \qquad z(t) \geq s(t),$$

given $z(0) > 0$.

This problem may be regarded as a production-planning problem with no inventory costs but with known demand $d(t) = \dot{s}(t)$ (see Chapter 1). Thus at any t we must have

$$z(0) + \int_0^t r(\tau)\, d\tau \geq \int_0^t d(\tau)\, d\tau$$

or, equivalently, $z(t) \geq s(t)$. We solve this problem for the specific case:

$$z(0) = \tfrac{1}{2}$$

$$s(t) = \begin{cases} 2t & 0 \leq t \leq \tfrac{1}{2} \\ 1 & \tfrac{1}{2} < t \leq 1. \end{cases}$$

Let us consider r as a member of a space X. The choice of X is somewhat free but we wish to choose it so that the corresponding constraint space Z can be taken to be $C[0, 1]$. In other words, we want a class of r's such that

$$z(t) = z(0) + \int_0^t r(\tau)\, d\tau$$

will be continuous. The choice $X = L_1[0, 1]$ is awkward because $\int_0^1 r^2\, dt$ does not exist for all $r \in L_1[0, 1]$. The choice $X = C[0, 1]$ is feasible but,

as we see from Example 3, there may be no solution in this space. Two other possibilities are: (1) the space of piecewise continuous functions and (2) the space of bounded measurable functions. We may define the supremum norm on these last two spaces but this is unnecessary since Lemma 1 is valid for linear Gateaux differentials as well as Fréchet differentials and this example goes through either way. Let us settle on piecewise continuous functions.

The Lagrangian for this problem is

$$L(r, v) = \tfrac{1}{2}\int_0^1 r^2(t)\,dt + \int_0^1 [s(t) - z(t)]\,dv(t),$$

where $v \in BV[0, 1]$ and is nondecreasing. We seek a saddle point $r_0 \geq \theta$, $v_0 \geq \theta$.

A more explicit formula for the Lagrangian is

$$\begin{aligned} L(r, v) &= \tfrac{1}{2}\int_0^1 r^2(t)\,dt + \int_0^1 [s(t) - z(0)]\,dv(t) - \int_0^1 \int_0^t r(\tau)\,d\tau\,dv(t) \\ &= \tfrac{1}{2}\int_0^1 r^2(t)\,dt + \int_0^1 [s(t) - z(0)]\,dv(t) \\ &\qquad + \int_0^1 r(t)v(t)\,dt - v(1)\int_0^1 r(t)\,dt. \end{aligned}$$

Using Lemma 1 to minimize $L(r, v_0)$ over $r \geq \theta$, we require that

$$r_0(t) + v_0(t) - v_0(1) \geq 0 \qquad \text{for all } t$$

$$r_0(t)[r_0(t) + v_0(t) - v_0(1)] = 0 \qquad \text{for all } t.$$

Since $s(t) \leq z(t)$ for all t, in order to maximize $L(r_0, v)$ we require that $v_0(t)$ varies only for those t with $z(t) = s(t)$.

The set of functions satisfying these requirements is shown in Figure 8.5. Note that v is a step function and varies only at the single point $t = \frac{1}{2}$. If a Lagrangian of the form

$$L(r, \lambda) = \tfrac{1}{2}\int_0^1 r^2(t)\,dt + \int_0^1 [s(t) - z(t)]\lambda(t)\,dt$$

were considered, no saddle point could be found unless functions λ having delta functions were allowed.

The economic interpretation of the Lagrange multiplier can be obtained rather directly. Suppose that the demand function were changed by $\Delta d(t)$. Letting ΔJ and Δs be the corresponding changes in the total production cost and in the function s, we have to first order

$$\Delta J = \int_0^1 \Delta s(t)\,dv(t) = -\int_0^1 \Delta s v(t)\,dt + \Delta s(1)v(1) - \Delta s(0)v(0).$$

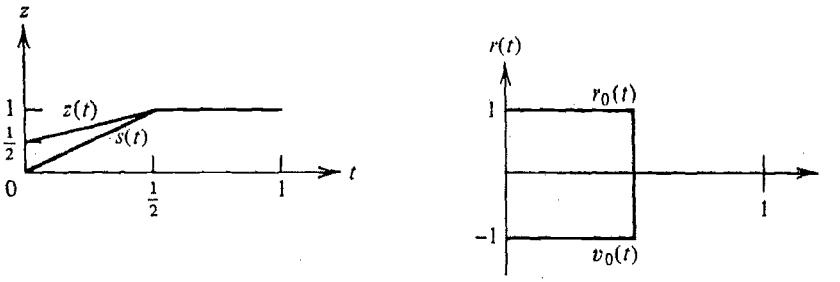

Figure 8.5 The solution to a production problem

Since $\Delta s(0) = 0$, $v(1) = 0$, the boundary terms vanish and hence

$$\Delta J = -\int_0^1 \Delta d(t)v(t)\,dt.$$

Thus $-v(t)$ is the price per unit that must be charged for small additional orders at time t to recover the increased production cost. Note that in the particular example considered above this price is zero for $t > \frac{1}{2}$. This is because the production cost has zero slope at $r = 0$.

8.8 Problems

1. Prove that the subset P of $L_p[a, b]$, $1 \leq p < \infty$, consisting of functions nonnegative almost everywhere on $[a, b]$, contains no interior point.
2. Prove that the subset P of $C[a, b]$ or $L_\infty[a, b]$, consisting of nonnegative functions on $[a, b]$, contains interior points.
3. Let P be the cone in $C[a, b]$ consisting of nonnegative (continuous) functions on $[a, b]$. Show that $P^\oplus$ consists of those functions v of bounded variation on $[a, b]$ that are nondecreasing.
4. Show that the positive cone $P^\oplus$ in X^* is closed.
5. Show that the sum of two convex mappings is also convex.
6. A cone is said to be *pointed* if it contains no one-dimensional subspace. If X is a vector space with positive cone P, show that $x \geq \theta$ and $x \leq \theta$ imply $x = \theta$ if and only if P is pointed.
7. Let $\mu_0 = \inf\{f(x) : x \in \Omega,\ G(x) \leq \theta,\ H(x) = \theta\}$, where all entities are defined as in Theorem 1, Section 8.3, with the addition that $H(x) = Ax + y_0$ (where A is linear) is a map of X into the finite-dimensional normed space Y. Assume that μ_0 is finite, that $\theta \in Y$ is an interior point of $\{y \in Y : H(x) = y \text{ for some } x \in \Omega\}$, and that there exists $x_1 \in \Omega$ such that $G(x_1) < \theta$, $H(x_1) = \theta$. Show that there is $z_0^* \geq \theta$, y_0^* such that

$$\mu_0 = \inf\{f(x) + \langle G(x), z_0^* \rangle + \langle H(x), y_0^* \rangle : x \in \Omega\}.$$

8. Let f be a functional defined on a normed space X. An element $x_0^* \in X^*$ is said to be a *subgradient* of f at x_0 if

$$f(x) - f(x_0) \geq \langle x - x_0, x_0^* \rangle$$

for all $x \in X$. Show that if f has a gradient at x_0, any subgradient is in fact equal to the gradient.

9. Show that in the sense of Theorem 1, Section 8.6, the dual to the linear programming problem:

$$\begin{aligned} &\text{minimize } b'x \\ &\text{subject to } Ax \geq c, \qquad x \geq \theta, \end{aligned}$$

where x is an n vector, b is an n vector, c is an m vector, and A is an $m \times n$ matrix is the problem:

$$\begin{aligned} &\text{maximize } c'\lambda \\ &\text{subject to } A'\lambda \leq b, \qquad \lambda \geq \theta. \end{aligned}$$

10. Derive the minimum norm duality theorem from the Lagrange duality theorem.
11. Derive necessary conditions for the problem: minimize $f(x)$, subject to $x \in \Omega$, $G(x) \leq z$, $z \in \Lambda$, where both x and z are variable. Assume that all functions and sets are convex.
12. Assuming that the interior point condition is satisfied, show that the dual of a convex-programming problem is also a convex-programming problem. Show that the dual of the dual, with variables restricted to X rather than X^{**}, is, in some sense, the primal problem.
13. Let G be a convex mapping from $\Omega \subset X$ into a normed space Z and assume that a positive cone having nonempty interior is defined in Z. Show that the two regularity conditions are equivalent:
    (a) There is an $x_1 \in \Omega$ such that $G(x_1) < \theta$.
    (b) For every $z^* \geq \theta$, $z^* \neq \theta$, there is an $x \in \Omega$ such that $\langle G(x), z^* \rangle < 0$.
14. If we let

$$\begin{aligned} z(t) &= \text{production rate} \\ d(t) &= \text{demand rate} \\ y(t) &= \text{inventory stock}, \end{aligned}$$

a simple production system is governed by the equation

$$y(t) = y(0) + \int_0^t [z(\tau) - d(\tau)]\, d\tau.$$

Find the production plan $z(t)$ for the interval $0 \leq t \leq T$ satisfying $z(t) \geq 0$, $y(t) \geq 0$, $0 \leq t \leq T$, and minimizing the sum of the production costs and inventory costs

$$J = \int_0^T [\tfrac{1}{2} z^2(t) + h \cdot y(t)]\, dt.$$

Assume that

$$d(t) \equiv 1, \qquad 2y(0)h > 1, \qquad \frac{1}{2h} + y(0) > T, \qquad y(0) < T.$$

Answer:

$$z(t) = \begin{cases} 0 & 0 \le t \le t_1 \\ h \cdot (t - t_1) & t_1 \le t \le T, \end{cases}$$

where

$$t_1 = T - \sqrt{\frac{2}{h}[T - y(0)]}.$$

15. Repeat Problem 14 for

$$d(t) \equiv 1, \qquad 2y(0)h > 1, \qquad \frac{1}{2h} + y(0) \le T, \qquad y(0) < T.$$

It is helpful to define

$$t_1 = y(0) - \frac{1}{2h}, \quad t_2 = y(0) + \frac{1}{2h}.$$

REFERENCES

§8.1–4. Lagrange multipliers for problems having inequality constraints were first treated explicitly by John [76] and Kuhn and Tucker [91]. The approach presented in this chapter, which requires no differentiability assumptions, was developed by Slater [142] and extended to infinite-dimensional spaces by Hurwicz [75]. In fact, our presentation closely follows Hurwicz. An excellent treatment of convex programming in finite-dimensional spaces is contained in Karlin [82]. For some interesting extensions, see Arrow, Hurwicz, and Uzawa [12]. A similar theory can be developed for quasi-convex programming. See Luenberger [100].

§8.5–6. The sensitivity property was observed by Everett [48]. The general duality theorem evolved from Dorn [42], Wolfe [155], Hanson [69], Huard [74], and Mangasarian [102]. Also see Ritter [125] and Rissanen [124].

§8.7. Much of the solution to the oil-drilling problem is based on an analysis by Karlin [83, pp. 210–214] who obtains a solution derived from the Neyman-Pearson Lemma. The farmer's allocation problem was devised by Phillip Wolfe and David Gale as an example of continuous linear programming. Leonard Berkovitz noted an easy solution using control theory (see Section 9.6). Problems such as the production problem were solved using convex programming theory by Lack [93].

§8.8. Problems 14 and 15, together with numerous other cases, are solved in Arrow, Karlin, and Scarf [13, especially pp. 37–40].

9
LOCAL THEORY OF CONSTRAINED OPTIMIZATION

9.1 Introduction

Historically, the local theory of Lagrange multipliers, stated in differential form, predates the global theory presented in Chapter 8 by almost a century. Its wider range of applicability and its general convenience for most problems continue to make the local theory the better known and most used of the two.

The general underlying principles of the two theories are substantially the same. In fact the Lagrange multiplier result for inequality constraints goes through almost identically for both the local and global theories. But there are some important differences, particularly for equality constraints, stemming from the fact that the local theory is based on approximations in the primal space X and hence auxiliary analysis is required to relate these approximations to the constraint space Z or to Z^*, the space in which the Lagrange multiplier is defined. For this reason, adjoint operators play a significant part in the development of the local theory since they enable us to transfer results in X^* back to Z^*.

For problems with equality constraints only, the nicest result available is the general Lagrange multiplier theorem established by means of an ingenious proof devised by Liusternik. This theorem, which is by far the deepest Lagrange multiplier theorem in this book, is proved in Section 9.3. The difficult analysis underlying the theorem, however, is contained in a generalized inverse function theorem proved in Section 9.2.

In Section 9.4 an analogous theorem is given for optimization problems subject only to inequality constraints. The proof of this result is similar to the proof of the global Lagrange multiplier theorem of Chapter 8.

Following these general theorems on Lagrange multipliers, there are two sections devoted to optimal control theory. To some extent, this topic can be treated as a simple application of the general Lagrange multiplier theory. The structure of control problems, however, is worthy of special

attention because additional results can be derived for this important class of problems and additional insight into Lagrange multipliers is obtained from their analysis.

LAGRANGE MULTIPLIER THEOREMS

9.2 Inverse Function Theorem

In this section we prove a rather deep generalization of the classical inverse function theorem that enables us to derive the generalized Lagrange multiplier theorem in the next section. The inverse function theorem is of considerable importance in its own right in analysis; some variations of it are discussed in the problems at the end of this chapter. The proof of the theorem is quite complex and may be omitted at first reading, but it is important to understand the statement of the theorem and, in particular, to be familiar with the notion of a regular point.

Definition. Let T be a continuously Fréchet differentiable transformation from an open set D in a Banach space X into a Banach space Y. If $x_0 \in D$ is such that $T'(x_0)$ maps X onto Y, the point x_0 is said to be a *regular point* of the transformation T.

Example 1. If T is a mapping from E^n into E^m, a point $x_0 \in E^n$ is a regular point if the Jacobian matrix of T has rank m.

Theorem 1. *(Generalized Inverse Function Theorem) Let x_0 be a regular point of a transformation T mapping the Banach space X into the Banach space Y. Then there is a neighborhood $N(y_0)$ of the point $y_0 = T(x_0)$ (i.e., a sphere centered at y_0) and a constant K such that the equation $T(x) = y$ has a solution for every $y \in N(y_0)$ and the solution satisfies $\|x - x_0\| \leq K\|y - y_0\|$.*

Proof. Let L_0 = nullspace of $T'(x_0)$. Since L_0 is closed, the quotient space X/L_0 is a Banach space. Define the operator A on this space by $A[x] = T'(x_0)x$, where $[x]$ denotes the class of elements equivalent to x modulo L_0. The operator A is well defined since equivalent elements x yield identical elements $y \in Y$. Furthermore, this operator is linear, continuous, one-to-one, and onto; hence, by the Banach inverse theorem, A has a continuous linear inverse.

Given $y \in Y$ sufficiently close to y_0, we construct a sequence of elements $\{L_n\}$ from X/L_0 and a corresponding sequence $\{g_n\}$ with $g_n \in L_n$ such that $x_0 + g_n$ converges to a solution of $T(x) = y$. For fixed $y \in Y$, let $g_0 = \theta \in L_0$ and define the sequences $\{L_n\}$ and $\{g_n\}$ recursively by

$$L_n - L_{n-1} = A^{-1}(y - T(x_0 + g_{n-1})), \tag{1}$$

and from the coset L_n select g_n such that

$$\|g_n - g_{n-1}\| \leq 2\|L_n - L_{n-1}\|$$

(which is possible since $\|L_n - L_{n-1}\| = \inf_{g \in L_n} \|g - g_{n-1}\|$). Rewriting (1) slightly, we have

$$L_n = A^{-1}(y - T(x_0 + g_{n-1}) + T'(x_0)g_{n-1})$$

and similarly

$$L_{n-1} = A^{-1}(y - T(x_0 + g_{n-2}) + T'(x_0)g_{n-2}).$$

Therefore,

$$L_n - L_{n-1} = -A^{-1}(T(x_0 + g_{n-1}) - T(x_0 + g_{n-2}) - T'(x_0)(g_{n-1} - g_{n-2})).$$

Define $g_t = tg_{n-1} + (1-t)g_{n-2}$. By the generalized mean value inequality (Proposition 2, Section 7.3) applied to the transformation

$$\Gamma(x) = -A^{-1}(T(x) - T'(x_0)x),$$

we obtain

$$(2) \quad \|L_n - L_{n-1}\| \leq \|A^{-1}\| \|g_{n-1} - g_{n-2}\| \sup_{0<t<1} \|T'(x_0 + g_t) - T'(x_0)\|.$$

Since T' is continuous at x_0, given $\varepsilon > 0$, there is an $r > 0$ such that $\|T'(x) - T'(x_0)\| < \varepsilon$ for $\|x - x_0\| < r$. Assuming $\|g_{n-1}\| < r$, $\|g_{n-2}\| < r$, we have $\|g_t\| < r$; therefore (2) implies that

$$\|L_n - L_{n-1}\| \leq \varepsilon \|A^{-1}\| \|g_{n-1} - g_{n-2}\|.$$

Furthermore,

$$\|g_n - g_{n-1}\| \leq 2\|L_n - L_{n-1}\| \leq 2\varepsilon \|A^{-1}\| \|g_{n-1} - g_{n-2}\|$$

and hence for sufficiently small ε

$$(3) \quad \|g_n - g_{n-1}\| \leq \tfrac{1}{2}\|g_{n-1} - g_{n-2}\|.$$

Since $\|g_1\| \leq 2\|L_1\| \leq 2\|A^{-1}\| \|y - y_0\|$, it follows that for $\|y - y_0\|$ sufficiently small we have

$$\|g_1\| < \tfrac{1}{2}r.$$

Thus the conditions for (3) to be valid hold for $n = 2$ and, in fact, hold for all n since, by induction, the relation

$$\begin{aligned} \|g_n\| &= \|g_1 + (g_2 - g_1) + \cdots + (g_n - g_{n-1})\| \\ &\leq \left(1 + \tfrac{1}{2} + \cdots + \frac{1}{2^{n-1}}\right)\|g_1\| \leq 2\|g_1\| \leq r \end{aligned}$$

shows that $\|g_n\| \leq r$ for all n.

Thus

$$\|g_n - g_{n-1}\| \leq \tfrac{1}{2}\|g_{n-1} - g_{n-2}\|$$

for all n and hence the sequence $\{g_n\}$ converges to an element g. Correspondingly, the sequence $\{L_n\}$ converges to a coset L. For these limits we have

$$L = L + A^{-1}(y - T(x_0 + g))$$

or, equivalently,

$$T(x_0 + g) = y;$$

and

$$\|g\| \leq 2\|g_1\| \leq 4\|A^{-1}\|\|y - y_0\|,$$

thus K can be taken to be $4\|A^{-1}\|$. ∎

9.3 Equality Constraints

Our aim now is to develop necessary conditions for an extremum of f subject to $H(x) = \theta$ where f is a real-valued functional on a Banach space X and H is a mapping from X into a Banach space Z.

Lemma 1. *Let f achieve a local extremum subject to $H(x) = \theta$ at the point x_0 and assume that f and H are continuously Fréchet differentiable in an open set containing x_0 and that x_0 is a regular point of H. Then $f'(x_0)h = 0$ for all h satisfying $H'(x_0)h = \theta$.*

Proof. To be specific, assume that the local extremum is a local minimum. Consider the transformation $T: X \to R \times Z$ defined by $T(x) = (f(x), H(x))$. If there were an h such that $H'(x_0)h = \theta, f'(x_0)h \neq 0$, then $T'(x_0) = (f'(x_0), H'(x_0)) : X \to R \times Z$ would be *onto* $R \times Z$ since $H'(x_0)$ is *onto* Z. By the inverse function theorem, it would follow that for any $\varepsilon > 0$ there exists a vector x and $\delta > 0$ with $\|x - x_0\| < \varepsilon$ such that $T(x) = (f(x_0) - \delta, \theta)$, contradicting the assumption that x_0 is a local minimum. ∎

The above result can be visualized geometrically in the space X in terms of the concept of the tangent space of the constraint surface. By the tangent space at x_0, we mean the set of all vectors h for which $H'(x_0)h = \theta$ (i.e., the nullspace of $H'(x_0)$). It is a subspace of X which, when translated to the point x_0, can be regarded as the tangent of the surface $N = \{x : H(x) = \theta\}$ near x_0. An equivalent statement to that of Lemma 1 is that f is stationary at x_0 with respect to variation in the tangent plane. This is illustrated in Figure 9.1 where contours of f as well as the constraint surface for a single

functional constraint $h(x) = 0$ are drawn. The Lagrange multiplier theorem now follows easily from the duality relations between the range and null-space of an operator and its adjoint.

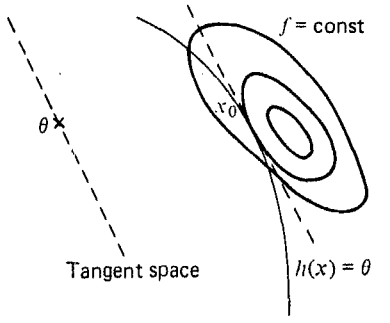

Figure 9.1 Constrained optimization

Theorem 1. (*Lagrange Multiplier*) *If the continuously Fréchet differentiable functional f has a local extremum under the constraint $H(x) = \theta$ at the regular point x_0, then there exists an element $z_0^* \in Z^*$ such that the Lagrangian functional*[1]

$$L(x) = f(x) + z_0^* H(x)$$

is stationary at x_0, i.e., $f'(x_0) + z_0^* H'(x_0) = \theta$.

Proof. From Lemma 1 it is clear that $f'(x_0)$ is orthogonal to the null-space of $H'(x_0)$. Since, however, the range of $H'(x_0)$ is closed, it follows (by Theorem 2, Section 6.6) that

$$f'(x_0) \in \mathscr{R}[H'(x_0)^*].$$

Hence there is a $z_0^* \in Z^*$ such that

$$f'(x_0) = -H'(x_0)^* z_0^*$$

or, in an alternative notation,

$$f'(x_0) + z_0^* H'(x_0) = \theta.$$

The second term of this last expression is to be interpreted in the usual sense of composition of linear transformations. ∎

We may immediately rephrase this result to include the case when the constraint is not regular.

[1] When considering the local theory the necessary conditions are written $f'(x_0) + z_0^* H'(x_0) = \theta$. Therefore we usually write the Lagrangian in the form indicated in the theorem statement rather than $L(x) = f(x) + \langle H(x), z_0^* \rangle$.

Corollary 1. *Assuming all the hypotheses of Theorem* 1 *with the exception that the range of* $H'(x_0)$ *is closed but perhaps not onto, there exists a nonzero element* $(r_0, z_0^*) \in R \times Z^*$ *such that the functional*

$$r_0 f(x) + z_0^* H(x)$$

is stationary at x_0.

Proof. If x_0 is regular, we may take $r_0 = 1$ and use Theorem 1. If x_0 is not a regular point, let $M = \mathscr{R}(H'(x_0))$. There is a point $z \in Z$ such that

$$\inf_{m \in M} \|z - m\| > 0.$$

and hence by Theorem 1, Section 5.8, there is a $z_0^* \in M^\perp$, $z_0^* \neq \theta$. Since by Theorem 1, Section 6.6, $M^\perp = \mathscr{N}(H'(x_0)^*)$, this z_0^* together with $r_0 = 0$ satisfies the requirements of the corollary. ∎

Reviewing the arguments that led to the Lagrange multiplier theorem, it should be noted that the difficult task, requiring the regularity assumption, was to show that at an extremum f is stationary with respect to variations in the tangent plane; in other words, justifying that a nonlinear constraint can be replaced by a linearized version of it. From this result it is clear that $f'(x_0)$, the gradient of f, must be orthogonal to the tangent space. The Lagrange multiplier theorem then follows from the familiar adjoint relations.

A slightly different but useful interpretation yields a more direct identification of the Lagrange multiplier. In Figure 9.1 the constraint is described by a single functional equation $h(x) = 0$; it is clear on geometric grounds that, at the optimum, the gradient of h must be parallel to the gradient of f. Thus $f'(x_0) + h'(x_0)z = \theta$, where z is a scalar: the Lagrange multiplier. A similar figure can be drawn in three dimensions for the case where H consists of two functionals h_1, h_2. An example is shown in Figure 9.2. For optimality the gradient of f must lie in the plane generated by h_1' and h_2'; hence $f'(x_0) + z_1 h_1'(x_0) + z_2 h_2'(x_0) = \theta$.

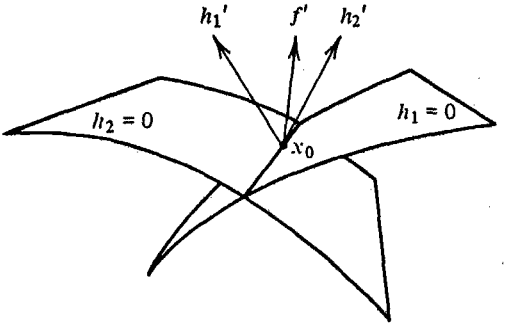

Figure 9.2 Two constraints

Example 1. (Constrained Problems in the Calculus of Variations) We consider the problem of finding $x \in D^n[t_0, t_1]$, the space of n-vector functions on $[t_0, t_1]$ possessing continuous derivatives, having fixed end points $x(t_0)$, $x(t_1)$, which minimizes

$$J = \int_{t_0}^{t_1} f(x, \dot{x}, t)\, dt \tag{1}$$

while satisfying the constraint

$$\phi(x, t) = 0. \tag{2}$$

Both f and ϕ are real-valued functions and are assumed to have continuous partial derivatives of second order. Since the end points are fixed, we restrict our attention to variations lying in the subspace $X \subset D^n[t_0, t_1]$ consisting of those functions that vanish at t_0 and t_1. As demonstrated in Section 7.5, the Fréchet differential of J is

$$\delta J(x; h) = \int_{t_0}^{t_1} [f_x(x, \dot{x}, t)h(t) + f_{\dot{x}}(x, \dot{x}, t)\dot{h}(t)]dt. \tag{3}$$

The function ϕ can be considered as a mapping H from X into Y where Y is the subspace of $D[t_0, t_1]$ consisting of those functions vanishing at t_0 and t_1. The Fréchet differential of H is

$$\delta H(x; h) = \phi_x h(t). \tag{4}$$

We assume that along the minimizing curve the n partial derivatives of ϕ (with respect to the n components of x) do not all simultaneously vanish at any point in $[t_0, t_1]$. In this case the Fréchet differential (4), evaluated at the minimizing x, defines a bounded linear mapping from X *onto* Y. To verify this we note that, since ϕ has continuous second partial derivatives, (4) defines an element of Y for each $h \in X$. Also, given $y \in Y$, the selection

$$h(t) = \frac{\phi_x' y(t)}{\phi_x \phi_x'}$$

(where ϕ_x is represented as an $n \times 1$ row vector and ϕ_x' is its transpose) satisfies $\phi_x h = y$. Thus, according to Theorem 1, there exists a Lagrange multiplier for the problem. The multiplier is in this case an element of Y^* which in general can be represented in the form

$$\langle y, y^* \rangle = \int_{t_0}^{t_1} \dot{y}(t)\, dz(t)$$

where $z \in NBV[t_0, t_1]$. However, it can be shown that the multiplier for this problem actually takes the special form

$$\langle y, y^* \rangle = \int_{t_0}^{t_1} y(t)\lambda(t)\, dt$$

for some continuous function λ. Hence the necessary conditions become

$$f_x(x, \dot{x}, t) + \lambda(t)\phi_x(x, \dot{x}, t) = \frac{d}{dt} f_{\dot{x}}(x, \dot{x}, t). \tag{5}$$

Example 2. (Geodesics) Let x, y, z denote the coordinates of an arbitrary point in three-dimensional space. The distance between two given points along a smooth arc $x(t), y(t), z(t)$ (parametrized by t, $t_1 \leq t \leq t_2$) is

$$J = \int_{t_1}^{t_2} \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}\, dt.$$

Given a smooth surface defined by $\phi(x, y, z) = 0$ and two points on this surface, the arc of minimum length lying in the surface $\phi(x, y, z) = 0$ and connecting the two points is called the *geodesic* between the points. If $|\phi_x| + |\phi_y| + |\phi_z| \neq 0$ along the geodesic, the method of Example 1 may be applied.

Writing the Lagrangian in the form

$$\int_{t_1}^{t_2} [\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2} + \lambda(t)\phi(x, y, z)]\, dt$$

the equations corresponding to equation (5) are

$$\frac{d}{dt} \frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}} + \lambda(t)\phi_x = 0$$

$$\frac{d}{dt} \frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}} + \lambda(t)\phi_y = 0$$

$$\frac{d}{dt} \frac{\dot{z}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}} + \lambda(t)\phi_z = 0$$

which, together with the constraint, can be solved for the arc.

In the special case of geodesics on a sphere, we have

$$\phi(x, y, z) = x^2 + y^2 + z^2 - r^2 = 0$$

and therefore

$$\phi_x = 2x, \qquad \phi_y = 2y, \qquad \phi_z = 2z.$$

Let the constants a, b, c be chosen so that the plane through the origin described by

$$ax + by + cz = 0$$

contains the two given points on the sphere.

Corresponding to the extremal arc $x(t)$, $y(t)$, $z(t)$, let $p(t) = ax(t) + by(t) + cz(t)$. Then it can be seen that p satisfies

$$\frac{d}{dt}\frac{\dot{p}}{\sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}} + 2\lambda(t)p(t) = 0$$

and $p(t_1) = p(t_2) = 0$. It follows from the uniqueness of solutions to differential equations, that $p(t) \equiv 0$ and hence that the geodesic lies in the plane. Therefore, the geodesic is a segment of a great circle on the sphere. The explicit dependence on t is, of course, somewhat arbitrary since any parametrization of the arc is allowed.

9.4 Inequality Constraints (Kuhn-Tucker Theorem)

In this section we derive the local necessary conditions for the problem

$$\text{(1)} \qquad \begin{cases} \text{minimize } f(x) \\ \text{subject to } G(x) \leq \theta, \end{cases}$$

where f is defined on a vector space X and G is a mapping from X into the normed space Z having positive cone P.

To see how the Lagrange multiplier technique can be extended to problems of this type, consider a problem in two dimensions having three scalar equations $g_i(x) \leq 0$ as constraints. Figure 9.3*a* shows the constraint region. In 9.3*b*, where it is assumed that the minimum occurs at a point x_0 in the interior of the region, it is clear that $f'(x_0) = 0$. In 9.3*c*, where it is assumed that the minimum occurs on the boundary $g_1(x) = 0$, it is clear that $f'(x_0)$ must be orthogonal to the boundary and point inside. Therefore, in this case, $f'(x_0) + \lambda_1 g_1'(x_0) = \theta$ for some $\lambda_1 \geq 0$. Similarly, in 9.3*d*, where it is assumed that the minimizing point x_0 satisfies both $g_1(x_0) = 0$ and $g_2(x_0) = 0$, we must have $f'(x_0) + \lambda_1 g_1'(x_0) + \lambda_2 g_2'(x_0) = \theta$ with $\lambda_1 \geq 0, \lambda_2 \geq 0$. All of these cases can be summarized by the general statement

$$f'(x_0) + \lambda^* G'(x_0) = 0,$$

where $\lambda^* \geq \theta$ and $\lambda_i g_i(x_0) = 0$, $i = 1, 2, 3$. The equality $\lambda_i g_i(x_0) = 0$ merely says that if $g_i(x_0) < 0$, then the corresponding Lagrange multiplier is absent from the necessary condition.

By considering various positive cones (such as $P = \{\theta\}$) or by considering constraints of the form $H(x) \leq \theta$, $-H(x) \leq \theta$, problem (1), as stated, includes minimization problems having equality constraints. It is therefore clear that a general attack on problem (1) would be at least as difficult to carry through as the corresponding attack on problems having only equality constraints. To avoid these difficulties, our approach excludes the possibility of equality constraints. As a consequence of this restriction, the results for problem (1), although analogous to those for problems having equality constraints, are far easier to obtain. This approach closely parallels the development of the global Lagrange multiplier theorem of Section 8.3.

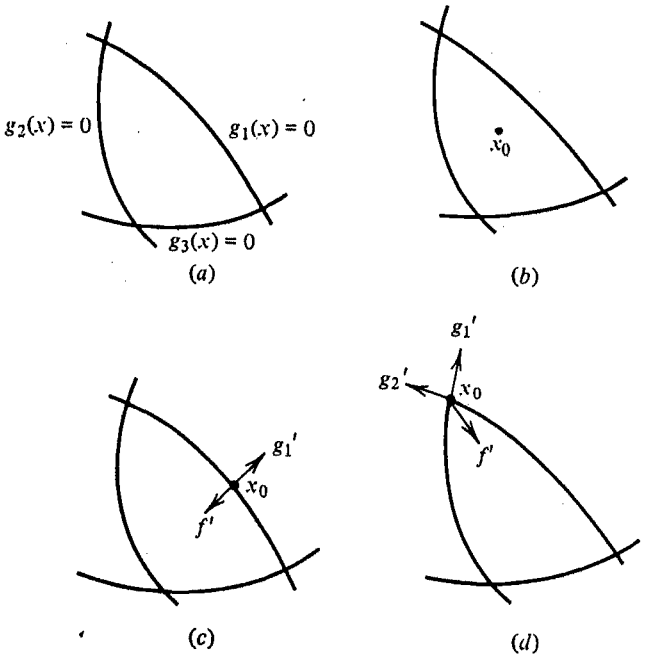

Figure 9.3 Inequality constraints

Definition. Let X be a vector space and let Z be a normed space with a positive cone P having nonempty interior. Let G be a mapping $G : X \to Z$ which has a Gateaux differential that is linear in its increment. A point $x_0 \in X$ is said to be a *regular point* of the inequality $G(x) \leq \theta$ if $G(x_0) \leq \theta$ and there is an $h \in X$ such that $G(x_0) + \delta G(x_0; h) < \theta$.

This definition of a regular point is a natural analog to the interior point condition employed for inequality constraints in the global theory (Theorem 1, Section 8.3). Note that the definition excludes the possibility

of incorporating equality constraints by reducing the cone to a point or by including a constraint and its negative.

The regularity condition essentially eliminates the possibility of the constraint boundary forming a cusp at a point. An example where the regularity condition fails is shown in Figure 9.4 where $g_1(x) = -x_2$,

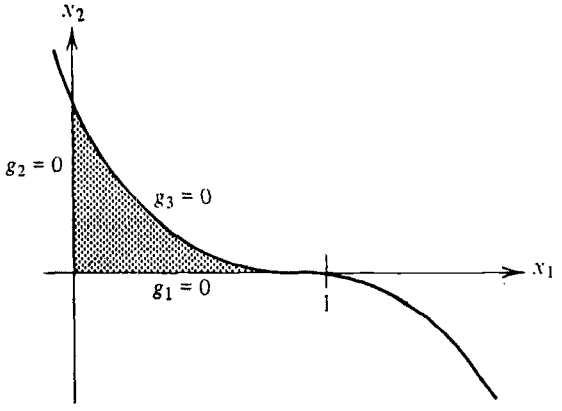

Figure 9.4 The regularity condition violated

$g_2(x) = -x_1$, $g_3(x) = x_2 + (x_1 - 1)^3$. The point $x_1 = 1$, $x_2 = 0$ is not regular since the gradients of g_2 and g_3 point in opposite directions there.

Theorem 1. (*Generalized Kuhn-Tucker Theorem*) *Let X be a vector space and Z a normed space having positive cone P. Assume that P contains an interior point.*

Let f be a Gateaux differentiable real-valued functional on X and G a Gateaux differentiable mapping from X into Z. Assume that the Gateaux differentials are linear in their increments.[1] *Suppose x_0 minimizes f subject to $G(x) \leq \theta$ and that x_0 is a regular point of the inequality $G(x) \leq \theta$. Then there is a $z_0^* \in Z^*$, $z_0^* \geq \theta$ such that the Lagrangian*

$$f(x) + \langle G(x), z_0^* \rangle$$

is stationary at x_0; furthermore, $\langle G(x_0), z_0^* \rangle = 0$.

Proof. In the space $W = R \times Z$, define the sets

$$A = \{(r, z) : r \geq \delta f(x_0; h),\ z \geq G(x_0) + \delta G(x_0; h) \text{ for some } h \in X\}$$

$$B = \{(r, z) : r \leq 0,\ z \leq \theta\}.$$

[1] As discusseed in Problem 9 at the end of this chapter, these hypotheses can be somewhat weakened.

The sets A and B are obviously convex; in fact, both are convex cones although A does not necessarily have its vertex at the origin. The set B contains interior points since P does.

The set A does not contain any interior points of B because if $(r, z) \in A$, with $r < 0$, $z < \theta$, then there exists $h \in X$ such that

$$\delta f(x_0; h) < 0, \qquad G(x_0) + \delta G(x_0; h) < \theta.$$

The point $G(x_0) + \delta G(x_0; h)$ is the center of some open sphere contained in the negative cone N in Z. Suppose this sphere has radius ρ. Then for $0 < \alpha < 1$ the point $\alpha[G(x_0) + \delta G(x_0; h)]$ is the center of an open sphere of radius $\alpha \cdot \rho$ contained in N; hence so is the point $(1 - \alpha)G(x_0) + \alpha[G(x_0) + \delta G(x_0; h)] = G(x_0) + \alpha \cdot \delta G(x_0; h)$. Since for fixed h

$$\|G(x_0 + \alpha h) - G(x_0) - \alpha \cdot \delta G(x_0; h)\| = o(\alpha),$$

it follows that for sufficiently small α, $G(x_0 + \alpha h) < \theta$. A similar argument shows that $f(x_0 + \alpha h) < f(x_0)$ for sufficiently small α. This contradicts the optimality of x_0; therefore A contains no interior points of B.

According to Theorem 3, Section 5.12, there is a closed hyperplane separating A and B. Hence there are r_0, z_0^*, δ such that

$$r_0 \cdot r + \langle z, z_0^* \rangle \geq \delta \qquad \text{for all } (r, z) \in A$$

$$r_0 \cdot r + \langle z, z_0^* \rangle \leq \delta \qquad \text{for all } (r, z) \in B.$$

Since $(0, \theta)$ is in both A and B, we have $\delta = 0$. From the nature of B it follows at once that $r_0 \geq 0$, $z_0^* \geq \theta$. Furthermore, the hyperplane cannot be vertical because of the existence of h such that $G(x_0) + \delta G(x_0; h) < \theta$. Therefore, we take $r_0 = 1$.

From the separation property, we have

$$\delta f(x_0; h) + \langle G(x_0) + \delta G(x_0; h), z_0^* \rangle \geq 0$$

for all $h \in X$. Setting $h = \theta$ gives $\langle G(x_0), z_0^* \rangle \geq 0$ but $G(x_0) \leq \theta$, $z_0^* \geq \theta$ implies $\langle G(x_0), z_0^* \rangle \leq 0$ and hence $\langle G(x_0), z_0^* \rangle = 0$. It then follows from the linearity of the differentials with respect to their increments that $\delta f(x_0; h) + \langle \delta G(x_0; h), z_0^* \rangle = 0$. ∎

Example 1. Suppose that X is a normed space (rather than simply a vector space) and that f and G are Fréchet differentiable. Then if the solution is at a regular point, we may write the conclusion of Theorem 1 as

$$f'(x_0) + z_0^* G'(x_0) = \theta$$

$$\langle G(x_0), z_0^* \rangle = 0.$$

Example 2. Consider the n-dimensional mathematical programming problem

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } G(x) \leq \theta \\ &\qquad\qquad x \geq \theta \end{aligned}$$

where $x \in E^n$, $G(x) \in E^m$, and f and G have continuous partial derivatives with respect to the components of x. The constraint for this problem can be written in partitioned form as

$$\begin{bmatrix} G(x) \\ -x \end{bmatrix} \leq \theta$$

which has E^{n+m} as the constraint space. Assuming satisfaction of the regularity condition, we have the existence of two Lagrange multiplier vectors $\lambda_0 \in E^m$, $\mu_0 \in E^n$ with $\lambda_0 \geq \theta$, $\mu_0 \geq \theta$ such that at the solution x_0

$$f_x(x_0) + \lambda_0' G_x(x_0) - \mu_0' = 0 \tag{1}$$

$$\lambda_0' G(x_0) - \mu_0' x_0 = 0. \tag{2}$$

Since the first term of (2) is nonpositive and the second term is nonnegative, they must both be zero. Thus, defining the reduced Lagrangian

$$L(x, \lambda) = f(x) + \lambda' G(x),$$

the necessary conditions can be written as

$$\begin{aligned} &L_x(x_0, \lambda_0) \geq \theta \qquad L_x(x_0, \lambda_0)x_0 = 0 \qquad x \geq \theta \\ &L_\lambda(x_0, \lambda_0) \leq \theta \qquad L_\lambda(x_0, \lambda_0)\lambda_0 = 0 \qquad \lambda \geq \theta. \end{aligned}$$

The top row of these equations says that the derivative of the Lagrangian with respect to x_i must vanish if $x_i > 0$ and must be nonnegative if $x_i = 0$. The bottom row says that λ_j is zero if the j-th constraint is not active, i.e., if the j-th component of G is not zero.

Example 3. Consider the constrained calculus of variations problem

$$\text{minimize } J = \int_{t_0}^{t_1} f(x, \dot{x}, t)\, dt \tag{3}$$

$$\text{subject to } \phi(x, t) \leq 0. \tag{4}$$

Here t_0, t_1 are fixed and x is a function of t. The initial value $x(t_0)$ is fixed and satisfies $\phi(x(t_0), t_0) < 0$. The real-valued functions f and ϕ have continuous partial derivatives with respect to their arguments. We seek a continuous solution $x(t)$ having piecewise continuous derivative. We assume that, along the solution, $\phi_x \neq 0$.

We incorporate the fact that $x(t_0)$ is fixed by restricting attention, in the variational analysis, to variations in the space X consisting of continuous functions vanishing at t_0 and having piecewise continuous derivatives on $[t_0, t_1]$. We consider the range of the constraint (4) to be $C[t_0, t_1]$. The regularity condition is then equivalent to the existence of an $h \in X$ such that

$$\phi(x_0(t), t) + \phi_x(x_0(t), t) \cdot h(t) < 0$$

for all $t \in [t_0, t_1]$. This condition is satisfied since it is assumed that $\phi(x_0(t_0), t_0) < 0$ and $\phi_x(x_0(t), t) \neq 0$.

We now obtain, directly from Theorem 1, the conditions

$$\int_{t_0}^{t_1} [f_x h(t) + f_{\dot{x}} \dot{h}(t)]\, dt + \int_{t_0}^{t_1} \phi_x h(t)\, d\lambda(t) = 0 \tag{5}$$

for all $h \in X$ with $h(t_0) = 0$, and

$$\int_{t_0}^{t_1} \phi(x, t)\, d\lambda = 0 \tag{6}$$

where $\lambda \in NBV[t_0, t_1]$ and is nondecreasing.

Integrating (5) by parts, we have

$$\int_{t_0}^{t_1} \left\{ -\int_{t_0}^{t} f_x\, d\tau + f_{\dot{x}} - \int_{t_0}^{t} \phi_x\, d\lambda \right\} h\, dt + h(t_1) \left\{ \int_{t_0}^{t_1} f_x\, dt + \int_{t_0}^{t_1} \phi_x\, d\lambda \right\} = 0. \tag{7}$$

The function

$$M(t) = -\int_{t_0}^{t} f_x\, d\tau + f_{\dot{x}} - \int_{t_0}^{t} \phi_x\, d\lambda$$

is bounded on $[t_0, t_1]$ and has at most a countable number of discontinuities. However, with the exception of the right end point, M must be continuous from the right. Therefore, by a slightly strengthened version of Lemma 2, Section 7.5, and by considering (7) for those particular $h \in X$ that vanish at t_1 as well as at t_0, we have

$$\int_{t_0}^{t} f_x\, d\tau + \int_{t_0}^{t} \phi_x\, d\lambda - f_{\dot{x}} = c \tag{8}$$

for $t \in [t_0, t_1)$. If λ does not have a jump at t_1, then (8) substituted into (7) yields

$$f_{\dot{x}}(x, \dot{x}, t)\Big|_{t=t_1} = 0$$

because $h(t_1)$ is arbitrary. On the other hand, if λ has a jump at t_1, then (8) implies that

$$\phi(x, t)\Big|_{t=t_1} = 0.$$

Therefore, we obtain in either case the boundary condition

$$\phi(x, t) \cdot f_{\dot{x}}(x, \dot{x}, t)\Big|_{t=t_1} = 0. \tag{9}$$

Together (6), (8), and (9) are a complete set of necessary conditions for the problem. For a simple application of this result, see Problem 6.

Example 4. As a specific instance of the above result, consider the problem

$$\text{minimize } J = \int_0^1 [x(t) + \tfrac{1}{2}\dot{x}(t)^2]\, dt$$

$$\text{subject to } x(t) \geq s(t).$$

Here s is a given continuous function and the initial condition $x(0) > s(0)$ is given.

Such a formulation might result from considering the problem of maintaining a work force x sufficient to handle a work level s when there is a linear salary cost and a quadratic cost for hiring and firing.

From (8) we obtain

$$t - \lambda(t) - \dot{x}(t) = c \tag{10}$$

where we have taken $\lambda(0) = 0$. Thus initially, while $x(t) > s(t)$, we have in view of equation (6)

$$\dot{x}(t) = \dot{x}(0) + t. \tag{11}$$

Let us hypothesize that the constraint $x(t) \geq s(t)$ is never achieved by equality. Then equation (11) must hold throughout [0, 1] and the terminal condition (9) implies in this case that $\dot{x}(1) = 0$. In other words, $x(t)$ is a parabola with second derivative equal to unity and having horizontal slope at the right end point.

If t_0 is a point where the solution meets the constraint, it is clear from equation (10) that the derivative must not have a positive jump at t_0; hence, unless $s(t_0)$ has a corner at t_0, x must be tangent to s at t_0. A typical solution is shown in Figure 9.5.

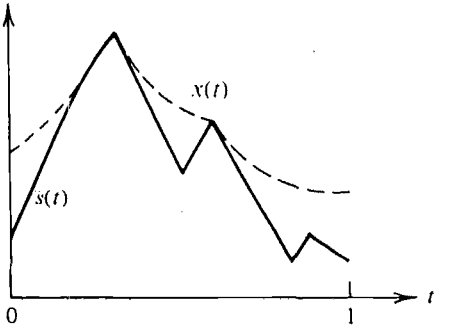

Figure 9.5 Solution to Example 4.

OPTIMAL CONTROL THEORY

9.5 Basic Necessary Conditions

On an interval $[t_0, t_1]$ of the real line, we consider a set of differential equations of the form

$$\dot{x}(t) = f(x(t), u(t)), \tag{1}$$

where $x(t)$ is an n-dimensional "state" vector, $u(t)$ is an m-dimensional "control" vector, and f is a mapping of $E^n \times E^m$ into E^n. Equation (1) describes a dynamic system which, when supplied with an initial state $x(t_0)$ and a control input function u, produces a vector-valued function x.

We assume that the vector-valued function f has continuous partial derivatives with respect to x and u. The class of admissible control functions is taken to be $C^m[t_0, t_1]$, the continuous m-dimensional functions on $[t_0, t_1]$, although there are other important alternatives. The space of admissible control functions is denoted U.

Given any $u \in U$ and an initial condition $x(t_0)$, we assume that equation (1) defines a unique continuous solution $x(t)$, $t > t_0$. The function x resulting from application of a given control u is said to be the trajectory of the system produced by u. The class of all admissible trajectories which we take to be the continuous n-dimensional functions on $[t_0, t_1]$ is denoted X.

In the classical optimal control problem, we are given, in addition to the dynamic equation (1) and the initial condition, an objective functional of the form

$$J = \int_{t_0}^{t_1} l(x, u)\, dt \tag{2}$$

and a finite number of terminal constraints

$$g_i(x(t_1)) = c_i \qquad i = 1, 2, \ldots, r$$

which we write in vector form as

$$G(x(t_1)) = c. \tag{3}$$

The functions l and G are assumed to possess continuous partial derivatives with respect to their arguments. The optimal control problem is then that of finding the pair of functions (x, u) minimizing J while satisfying (1) and (3).

There are a number of generalizations of this problem, many of which can be reduced to this form by appropriate transformations. For example,

problems in which the objective contains a function of the terminal state or problems in which the trajectory is constrained to satisfy a finite number of relations of the form

$$\int_{t_0}^{t_1} k(x, u)\, dt = d$$

can be transformed into the form considered above by adjoining additional state variables (see Problem 13). Problems in which the control variables are restricted by inequalities such as $|u(t)| \leq 1$ are discussed in Section 9.6.

When attempting to abstract the control problem so that the general variational theory can be applied, we discover several alternative viewpoints. Perhaps the most natural approach is to consider the problem as one formulated in $X \times U$ and to treat the differential equation (1) and the terminal constraint (3) as constraints connecting u and x; we then apply the general Lagrange multiplier theorem to these constraints. Another approach, however, is to note that (1) uniquely determines x once u is specified and hence we really only need to select u. The problem can thus be regarded as formulated in U; the Lagrange multiplier theorem need only be applied to the terminal constraints. Still another approach is to view the problem in X by considering the implicitly defined set of all trajectories that can be obtained by application of admissible controls. Finally, in Section 10.10 it is seen that it is sometimes profitable to view the problem in E^r, the finite-dimensional space corresponding to the constraint (3). Each of these approaches has theoretical advantages for the purpose of deriving necessary conditions and practical advantages for the purpose of developing computational procedures for obtaining solutions. In this section we approach the problem in the space $X \times U$ and in the next section in the space U.

The differential equation (1) with initial condition $x(t_0)$ is equivalent to the integral equation

$$x(t) - x(t_0) - \int_{t_0}^{t} f(x(\tau), u(\tau))\, d\tau = \theta \tag{4}$$

which we write abstractly as

$$A(x, u) = \theta. \tag{5}$$

The transformation A is a mapping from $X \times U$ into X. If we take $X = C^n[t_0, t_1]$, $U = C^m[t_0, t_1]$, then the Fréchet differential of A exists, is continuous under our assumptions, and is given by the formula

$$\delta A(x, u; h, v) = h(t) - \int_{t_0}^{t} f_x h(\tau)\, d\tau - \int_{t_0}^{t} f_u v(\tau)\, d\tau \tag{6}$$

for $(h, v) \in X \times U$.

The terminal constraint (3) is a mapping from X into E^r with Fréchet differential

$$\delta G(x; h) = G_x h(t_1). \tag{7}$$

Together the transformations A and G define the constraints of the problem, and we must investigate the question of regularity of these constraints. We must ask if, at the optimal trajectory, the Fréchet differentials (6) and (7) taken as a pair map *onto* $X \times E^r$ as (h, v) varies over $X \times U$.

From the differential (7) it is immediately clear that we must assume that the $r \times n$ matrix $G_x(x(t_1))$ has rank r. In addition we invoke a *controllability* assumption on (6). Specifically we assume that for any n-dimensional vector e it is possible to select a continuous function v such that the equation

$$h(t) - \int_{t_0}^{t} f_x h(\tau)\, d\tau - \int_{t_0}^{t} f_u v(\tau)\, d\tau = 0$$

has solution h with $h(t_1) = e$. An intuitive way of describing this assumption is to say that the original system (1), linearized about the optimal trajectory, can be driven from the origin to any point in E^n.

With the above two assumptions we can show that the constraints are regular. For this it is sufficient to show that for any $e \in E^n$ and any function $y \in X$ there is an $(h, v) \in X \times U$ such that

$$h(t) - \int_{t_0}^{t} f_x h(\tau)\, d\tau - \int_{t_0}^{t} f_u v(\tau)\, d\tau = y(t) \tag{8}$$

$$h(t_1) = e. \tag{9}$$

First, for $v = 0$ in equation (8), there is, by the fundamental existence theorem for linear Volterra integral equations (see Example 3, Section 10.2), a solution $\bar{h}$ of (8). We may then write (8) as

$$w(t) - \int_{t_0}^{t} f_x w(\tau)\, d\tau - \int_{t_0}^{t} f_u v(\tau)\, d\tau = 0 \tag{10}$$

where $w(t) = h(t) - \bar{h}(t)$. By the controllability assumption there is a v such that the solution to (10) has $w(t_1) = e - \bar{h}(t_1)$. Then $h(t) = w(t) + \bar{h}(t)$ is the desired solution to equations (8) and (9).

Having examined the question of regularity, we now give the basic necessary conditions satisfied by the solution to the optimal control problem.

Theorem 1. *Let* x_0, u_0 *minimize*

$$J = \int_{t_0}^{t_1} l(x, u)\, dt \tag{2}$$

subject to

$$\dot{x}(t) = f(x, u), \qquad x(t_0) \text{ fixed} \tag{1}$$

$$G(x(t_1)) = c \tag{3}$$

and assume that the regularity conditions are satisfied. Then there is an n-dimensional vector-valued function $\lambda(t)$ *and an r-dimensional vector* μ *such that for all* $t \in [t_0, t_1]$

$$-\dot{\lambda}(t) = [f_x'(x_0(t), u_0(t))]\lambda(t) + l_x'(x_0(t), u_0(t)) \tag{11}$$

$$\lambda(t_1) = G_x'(x_0(t_1))\mu \tag{12}$$

$$\lambda'(t) f_u(x_0(t), u_0(t)) + l_u(x_0(t), u_0(t)) = \theta. \tag{13}$$

Proof. The Lagrange multiplier theorem (Theorem 1, Section 9.3) yields immediately the existence of $\lambda \in NBV^n[t_0, t_1]$, $\mu \in E^r$ such that

$$\int_{t_0}^{t_1} l_x(x_0, u_0)h(t)\,dt + \int_{t_0}^{t_1} d\lambda'(t)\left[h(t) - \int_{t_0}^{t} f_x(x_0, u_0)h(\tau)\,d\tau\right] + \mu' G_x(x_0(t_1))h(t_1) = 0 \tag{14}$$

$$\int_{t_0}^{t_1} l_u(x_0, u_0)v(t)\,dt - \int_{t_0}^{t_1} d\lambda'(t) \int_{t_0}^{t} f_u(x_0, u_0)v(\tau)\,d\tau = 0 \tag{15}$$

for all $(h, v) \in X \times U$. Without loss of generality, we may take $\lambda(t_1) = \theta$.

Integrating (14) by parts, we have

$$\int_{t_0}^{t_1} l_x(x_0, u_0)h(t)\,dt + \int_{t_0}^{t_1} d\lambda'(t)h(t) + \int_{t_0}^{t_1} \lambda'(t) f_x(x_0, u_0)h(t)\,dt + \mu' G_x h(t_1) = 0. \tag{16}$$

It is clear that λ can have no jumps in $[t_0, t_1)$ since otherwise a suitable h could be constructed to make the second term of (16) large compared with the other terms. There must, however, be a jump at t_1 of magnitude $-G_x(x_0(t_1))\mu$. Since (16) holds for all continuous h, it holds in particular for all continuously differentiable h vanishing at t_0. Therefore, integrating the second term by parts, we have for such functions

$$\int_{t_0}^{t_1} \{l_x(x_0, u_0)h(t) - \lambda'(t)\dot{h}(t) + \lambda'(t) f_x(x_0, u_0)h(t)\}\,dt = 0.$$

Hence, by Lemma 3, Section 7.5, it follows that λ is differentiable on $[t_0, t_1)$ and that (11) holds.

Integrating equation (15) by parts, (13) follows from Lemma 1, Section 7.5. Now by changing the boundary condition on $\lambda(t_1)$ from $\lambda(t_1) = \theta$ to $\lambda(t_1) = G_x'\mu$ to account for the jump, λ will be continuous throughout $[t_0, t_1]$. ∎

Note that the conditions (11), (12), and (13) together with the original constraints (1) and (3) and the initial condition form a complete system of equations: $2n$ first-order differential equations, $2n$ boundary conditions, r terminal constraints, and m instantaneous equations from which $x_0(t)$, $\lambda(t)$, μ, and $u_0(t)$ can be found.

Example 1. We now find the m-dimensional control function u that minimizes the quadratic objective functional

$$J = \tfrac{1}{2}\int_{t_0}^{t_1} [x'(t)Qx(t) + u'(t)Ru(t)]\,dt \tag{17}$$

subject to the linear dynamic constraint

$$\dot{x}(t) = Fx(t) + Bu(t), \qquad x(t_0) \text{ fixed}, \tag{18}$$

where Q is an $n \times n$ symmetric positive-semidefinite matrix, R is an $m \times m$ symmetric positive-definite matrix, F is an $n \times n$ matrix, and B is an $n \times m$ matrix. This problem is of considerable importance in optimal control theory because it is sufficiently general to describe many practical problems adequately and is one of the few problems that can be solved explicitly.

Applying the necessary conditions of Theorem 1, we have

$$-\dot{\lambda}(t) = F'\lambda(t) + Qx(t), \qquad \lambda(t_1) = \theta \tag{19}$$

$$\lambda'(t)B + u'(t)R = 0. \tag{20}$$

Since R is positive definite, we have

$$u(t) = -R^{-1}B'\lambda(t). \tag{21}$$

Substituting equation (21) into (18), we obtain

$$\dot{x}(t) = Fx(t) - BR^{-1}B'\lambda(t), \qquad x(t_0) \text{ fixed}. \tag{22}$$

Together (19) and (22) form a linear system of differential equations in the variables x and λ. The system is complicated, however, by the fact that half of the boundary conditions are given at each end. To solve this system, we observe that the solution satisfies the relation

$$\lambda(t) = P(t)x(t), \tag{23}$$

where $P(t)$ is the $n \times n$ matrix solution of the Riccati differential equation

$$\dot{P}(t) = -P(t)F - F'P(t) + P(t)BR^{-1}B'P(t) - Q, \qquad P(t_1) = 0. \tag{24}$$

The verification follows by direct substitution and is left to the reader. It can be shown that (24) has a unique, symmetric, positive semidefinite solution on $[t_0, t_1]$.

From equations (21) and (23) we then have the solution

$$u(t) = -R^{-1}B'P(t)x(t) \tag{25}$$

which gives the control input in feedback form as a linear function of the state.

This solution is of great practical utility because if the solution $P(t)$ of equation (24) is found (as, for example, by simple backward numerical integration), the optimal control can be calculated in real time from physical measurements of $x(t)$.

Example 2. A rocket is to be launched from a point at time $t = 0$ with fixed initial velocity and direction. The rocket is propelled by thrust developed by the rocket motor and is acted upon by a uniform gravitational field and negligible atmospheric resistance. Given the motor thrust, we seek the thrust direction program that maximizes the range of the rocket on a horizontal plane.

The problem is sketched in Figure 9.6. Note that the final time T is determined by the impact on the horizontal plane and is an unknown variable.

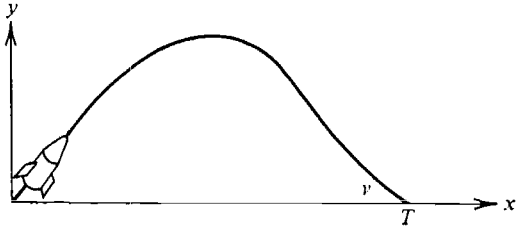

Figure 9.6 The rocket example

Letting $v_1 = \dot{x}$, $v_2 = \dot{y}$, the equations of motions are

$$\dot{v}_1 = \Gamma(t)\cos\theta, \qquad v_1(0) \text{ given},$$

$$\dot{v}_2 = \Gamma(t)\sin\theta - g, \qquad v_2(0) \text{ given},$$

where $\Gamma(t)$ is the instantaneous ratio of rocket thrust to mass and g is the acceleration due to gravity. The range is

$$J = \int_0^T v_1(t)\,dt = x(T),$$

where $T > 0$ is the time at which $y(T) = 0$ or, equivalently, the time at which $\int_0^T v_2(t)\,dt = 0$.

We may regard the problem as formulated in the space $C^2[0, T_0]$ of two-dimensional continuous time functions where T_0 is some fixed time greater

than the impact time of the optimal trajectory. Assuming a continuous nominal trajectory, $\bar{v}_1$, $\bar{v}_2$ with impact time $\overline{T}$, we can compute the Fréchet differential of J by reference to Figure 9.7 which shows the continuous nominal and a perturbed trajectory $v_1 = \bar{v}_1 + h_1$, $v_2 = \bar{v}_2 + h_2$. The perturbed trajectory crosses the x axis at a different time T. Denoting by $x(\overline{T})$, $y(\overline{T})$ the x and y coordinates of the perturbed trajectory at time $\overline{T}$, we have, to first order,

$$\overline{T} - T = \frac{y(\overline{T})}{\bar{v}_2(\overline{T})} = \frac{1}{\bar{v}_2(\overline{T})} \int_0^T h_2(t)\, dt$$

and

$$J(v + h) - J(v) = x(\overline{T}) - \bar{x}(\overline{T}) + (T - \overline{T})\bar{v}_1(\overline{T}) = \int_0^T h_1(t)\, dt + (T - \overline{T})\bar{v}_1(\overline{T}).$$

Combining these we have, to first order,

$$\delta J(\bar{v}; h) = J(\bar{v} + h) - J(\bar{v}) = \int_0^T h_1(t)\, dt - \frac{\bar{v}_1(\overline{T})}{\bar{v}_2(\overline{T})} \int_0^T h_2(t)\, dt.$$

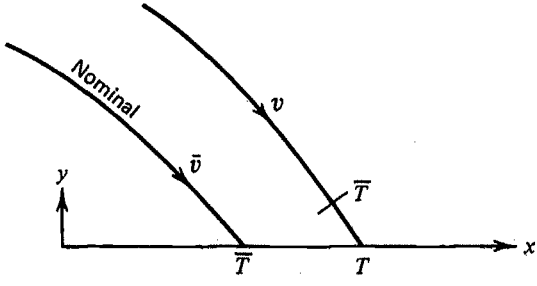

Figure 9.7 Calculation of Fréchet differential

Therefore, the original problem is equivalent to finding a stationary point of the functional

$$\int_0^T \left[v_1(t) - \frac{\bar{v}_1(\overline{T})}{\bar{v}_2(\overline{T})} v_2(t) \right] dt$$

which is an integral on the fixed interval $[0, \overline{T}]$. This problem can be solved in the standard fashion.

Following Theorem 1, we introduce the Lagrange variables

$$\lambda_1 = -1 \qquad \lambda_1(\overline{T}) = 0$$

$$\lambda_2 = \frac{\bar{v}_1(\overline{T})}{\bar{v}_2(\overline{T})} \qquad \lambda_2(\overline{T}) = 0.$$

Then $\lambda_1(t) = (\overline{T} - t), \lambda_2(t) = [v_1(\overline{T})/v_2(\overline{T})](t - \overline{T})$. The optimal $\theta(t)$ satisfies

$$-\lambda_1(t)\Gamma(t) \sin \theta(t) + \lambda_2(t)\Gamma(t) \cos \theta(t) = 0.$$

Thus we obtain the equation

$$\tan \theta(t) = -\frac{\bar{v}_1(\overline{T})}{\bar{v}_2(\overline{T})}.$$

We conclude that θ is constant in time. The constant is determined implicitly by the particular nature of $\Gamma(t)$.

*9.6 The Pontryagin Maximum Principle

The Pontryagin maximum principle gives a set of necessary conditions for control problems in which the control $u(t)$ is constrained to a given set. In this section we develop one form of the maximum principle from an abstract viewpoint by exploiting the basic definition of an adjoint operator.

Motivated by the framework of the optimal control problem discussed in Section 9.5, we let X and U be normed linear spaces and $g[x, u]$ a cost functional on $X \times U$, and we consider a constraint of the form

$$A[x, u] = \theta \tag{1}$$

where A is a mapping from $X \times U$ into X. The transformation A describes the system equations and may represent a set of differential equations, integral equations, partial differential equations, difference equations, etc. We assume that (1) defines a unique implicit function $x(u)$. Furthermore, we assume that A and g are Fréchet differentiable with respect to x and that the derivatives $A_x[x, u]$ and $g_x[x, u]$ are continuous on $X \times U$. Finally, we assume that the implicit function $x(u)$ satisfies a Lipschitz condition of the form

$$\|x(u) - x(v)\| \leq K\|u - v\|. \tag{2}$$

The control problem is to find (x, u) minimizing $J = g[x, u]$ while satisfying $A[x, u] = \theta$ and $u \in \Omega$ where Ω is a prescribed subset of U. Since x is uniquely determined from u, the objective functional J can be considered to be dependent only on u, it being understood that $J(u) = g[x(u), u]$.

We now introduce the Lagrangian functional of our constrained optimization problem. For $x \in X$, $u \in U$, $\lambda^* \in X^*$, we define

$$L[x, u, \lambda^*] = \lambda^* A[x, u] + g[x, u]. \tag{3}$$

The following proposition can be regarded as the basis of a number of necessary conditions for control problems connected with various types of systems.

Proposition 1. *For any $u \in \Omega$ let λ^* be a solution of the equation*

$$\lambda^* A_x[x(u), u] + g_x[x(u), u] = \theta. \tag{4}$$

Then for $v \in \Omega$,

$$J(u) - J(v) = L[x(u), u, \lambda^*] - L[x(u), v, \lambda^*] + o(\|u - v\|). \tag{5}$$

Proof. By definition

$$\begin{aligned} J(u) - J(v) &= g[x(u), u] - g[x(v), v] \\ &= g[x(u), u] - g[x(u), v] + g[x(u), v] - g[x(v), v] \\ &= g[x(u), u] - g[x(u), v] + g_x[x(u), u] \cdot [x(u) - x(v)] \\ &\quad + (g_x[x(v), v] - g_x[x(u), u])[x(u) - x(v)] + o(\|x(u) - x(v)\|) \\ &= g[x(u), u] - g[x(u), v] + g_x[x(u), u][x(u) - x(v)] \\ &\quad + o(\|u - v\|), \end{aligned}$$

where the last two steps follow from the continuity of g_x and the Lipschitz condition (2).

Likewise,

$$\|A[x(u), u] - A[x(u), v] - A_x[x(u), u][x(v) - x(u)]\| = o(\|v - u\|).$$

Therefore,

$$J(u) - J(v) = L[x(u), u, \lambda^*] - L[x(u), v, \lambda^*] + o(\|u - v\|). \quad \blacksquare$$

The significance of the above result is that it gives, to first order, a way of determining the change in J due to a change in u without reevaluating the implicit function x. The essence of the argument is brought out in Problem 18. We next apply this result to a system described by a system of ordinary differential equations of the form

$$\dot{x} = f(x, u), \qquad x(t_0) = x_0$$

and an associated objective functional

$$J = \int_{t_0}^{t_1} l(x, u)\, dt.$$

It is assumed that the functions f and l are continuously differentiable with respect to x and that f satisfies a uniform Lipschitz condition with respect to x and u of the form

$$\|f(x, u) - f(y, v)\| \leq M[\|x - y\| + \|u - v\|],$$

where $\|\ \|$ denotes the finite-dimensional norm.

Unlike in the previous section, we now take the admissible control functions u to be the piecewise continuous functions on the interval $[t_0, t_1]$, and require that for each t, $u(t) \in \Omega$ where Ω is a prescribed subset of E^m.

Theorem 1. *Let x_0, u_0 be optimal for the problem of minimizing*

$$J = \int_{t_0}^{t_1} l(x, u)\, dt$$

subject to $\dot{x}(t) = f(x, u)$, $x(t_0)$ fixed, $u(t) \in \Omega$. Let λ be the solution of the equation

$$-\dot{\lambda}(t) = f_x'\lambda(t) + l_x', \qquad \lambda(t_1) = 0, \tag{6}$$

where the partial derivatives are evaluated along the optimal trajectory, and define the Hamiltonian function

$$H(x, u, \lambda, t) = \lambda'(t) f(x, u) + l(x, u). \tag{7}$$

Then for all $t \in [t_0, t_1]$,

$$H(x_0(t), u_0(t), \lambda(t)) \leq H(x_0(t), u, \lambda(t)) \tag{8}$$

for all $u \in \Omega$.

Proof. For notational simplicity we assume $m = 1$; i.e., the controls are scalar functions. We take $X = C^n[t_0, t_1]$, and for U we take the space of piecewise continuous functions with the L_1 norm. Defining

$$A[x, u] = x(t) - x(t_0) - \int_{t_0}^{t} f(x(\tau), u(\tau))\, d\tau$$

$$g[x, u] = \int_{t_0}^{t_1} l(x, u)\, dt,$$

we see that A and g are continuously Fréchet differentiable with respect to x (although not with respect to u with the norm we are using).

If $x, x + \delta x$ correspond to $u, u + \delta u$, respectively, in $\{(x, u) : A[x, u] = \theta\}$, we have

$$\|\delta x(t)\|_{E^n} \leq \int_{t_0}^{t} M\{\|\delta x(\tau)\|_{E^n} + |\delta u(\tau)|\}\, d\tau$$

from which it follows that

$$\|\delta x(t)\|_{E^n} \leq M e^{M(t_1 - t_0)} \int_{t_0}^{t_1} |\delta u(\tau)|\, d\tau.$$

Therefore, $\|\delta x\| \leq K\|\delta u\|$ and the transformation $A[x, u]$ satisfies the Lipschitz condition required of Proposition 1.

It is clear that (6) is equivalent to the adjoint equation $\lambda^* A_x[x, u] + g_x[x, u] = \theta$. The functional

$$\int_{t_0}^{t_1} H(x, u, \lambda)\, dt$$

is then identical with the Lagrangian (3) except for a term $\int_{t_0}^{t_1} x'(t)\dot{\lambda}(t)\, dt$, which is not important since it does not depend explicitly on u. Proposition 1 gives us

$$(9) \quad J(u_0) - J(u) = \int_{t_0}^{t_1} [H(x_0, u_0, \lambda) - H(x_0, u, \lambda)]\, dt + o(\|u - u_0\|).$$

We now show that equation (9) implies (8). Suppose to the contrary that there is $\bar{t} \in [t_0, t_1]$ and $\bar{u} \in \Omega$ such that

$$H(x_0(\bar{t}), u_0(\bar{t}), \lambda(\bar{t})) > H(x_0(\bar{t}), \bar{u}, \lambda(\bar{t})).$$

In view of the piecewise continuity of u and the continuity of x, λ, f, and l, it follows that there is an interval $[t', t'']$ containing $\bar{t}$ and an $\varepsilon > 0$ such that

$$H(x_0(t), u_0(t), \lambda(t)) - H(x_0(t), \bar{u}, \lambda(t)) > \varepsilon$$

for all $t \in [t', t'']$.

Now let $u(t)$ be the piecewise continuous function equal to $u_0(t)$ outside $[t', t'']$ and equal to $\bar{u}$ on $[t', t'']$. From (9) we have

$$J(u_0) - J(u) > \varepsilon(t'' - t') + o(\|u - u_0\|).$$

But $\|u - u_0\| = 0([t'' - t'])$; hence, by selecting $[t', t'']$ sufficiently small, $J(u_0) - J(u)$ can be made positive, which contradicts the optimality of u_0. ∎

Before considering an example, several remarks concerning this result and its relation to other sets of necessary conditions are appropriate. Briefly, the result says that if a control function minimizes the objective functional, its values at each instant must also minimize the Hamiltonian. (Pontryagin's adjoint equation is defined slightly differently than ours with the result that his Hamiltonian must be maximized, thus accounting for the name maximum principle rather than minimum principle.) It should immediately be obvious that if the Hamiltonian is differentiable with respect to u as well as x and if the region Ω is open, the conditions of Theorem 1 are identical with those of Theorem 1, Section 9.5. The maximum principle can also be extended to problems having terminal constraints, but the proof is by no means elementary. However, problems of this type arising from applications are often convex and can be treated by the global theory developed in Chapter 8.

Example 1. We now solve the farmer's allocation problem (Example 3, Section 8.7) by the maximum principle. To formulate the problem as one of optimal control, we let $u(t)$ denote the fraction of production rate that is reinvested at time t. The problem then is described by

$$\dot{x}(t) = u(t)x(t), \qquad x(0) > 0 \tag{10}$$

$$J = \int_0^T (1 - u(t))x(t)\,dt \tag{11}$$

$$0 \le u(t) \le 1. \tag{12}$$

Here, as in Chapter 8, the farmer wishes to select u so as to maximize the total storage J.

The adjoint equation for this problem is

$$-\dot{\lambda}(t) = u(t)\lambda(t) + 1 - u(t), \qquad \lambda(T) = 0 \tag{13}$$

and the Hamiltonian is

$$H(x, u, \lambda) = \lambda(t)u(t)x(t) + [1 - u(t)]x(t). \tag{14}$$

An optimal solution x_0, u_0, λ must satisfy (10), (12), and (13) and (14) must be maximized with respect to admissible u's. Since $x(t) \ge 0$ for all $t \in [0, T]$, it follows from (14) that

$$u_0(t) = \begin{cases} 1 & \lambda(t) > 1 \\ 0 & \lambda(t) < 1. \end{cases}$$

Then since $\lambda(T) = 0$, we have $u_0(T) = 0$, and equation (13) can be integrated backward from $t = T$. The solution is shown in Figure 9.8. We conclude that the farmer stores nothing until $T - 1$, at which point he stores all production.

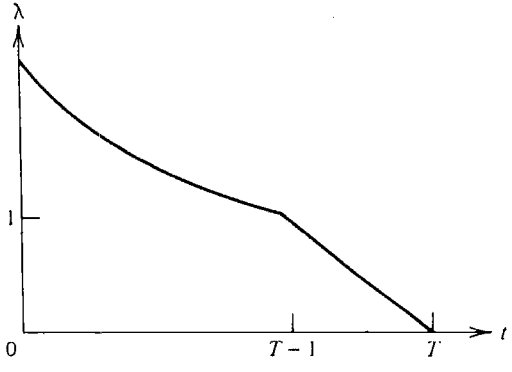

Figure 9.8 Solution to adjoint equation

9.7 Problems

1. Prove the following inverse function theorem. Let D be an open subset of a Banach space X and let T be a transformation from D into X. Assume that T is continuously Fréchet differentiable on D and that at a point $x_0 \in D$, $[T'(x_0)]^{-1}$ exists. Then:
   (i) There is a neighborhood P of x_0 such that T is one-to-one on P.
   (ii) There is a continuous transformation F defined on a neighborhood N of $T(x_0)$ with range $R \subset P$ such that $F(T(x)) = x$ for all $x \in R$.
   Hint: To prove uniqueness of solution to $T(x) = y$ in P, apply the mean value inequality to the transformation $\varphi(x) = [T'(x_0)]^{-1}T(x) - x$.
2. Prove the following implicit function theorem. Let X and Y be Banach spaces and let T be a continuously Fréchet differentiable transformation from an open set D in $X \times Y$ with values in X. Let (x_0, y_0) be a point in D for which $T(x_0, y_0) = \theta$ and for which $[T_x'(x_0, y_0)]^{-1}$ exists. Then there is a neighborhood N of y_0 and a continuous transformation F mapping N into X such that $F(y_0) = x_0$ and $T(F(y), y) = \theta$ for all $y \in N$.
3. Show that if all the hypotheses of Theorem 1, Section 9.4, are satisfied, except perhaps the regularity condition, then there is a nonzero, positive element $(r_0, z_0^*) \in R \times Z^*$ such that $r_0 f(x) + \langle G(x), z_0^* \rangle$ is stationary at x_0, and $\langle G(x), z_0^* \rangle = 0$.
4. Let $g_1, g_2, \ldots, g_n$ be real-valued Fréchet differentiable functionals on a normed space X. Let x_0 be a point in X satisfying

   $$g_i(x_0) \leq 0 \qquad i = 1, 2, \ldots, n. \tag{1}$$

   Let I be the set of indices i for which $g_i(x_0) = 0$ (the so-called binding constraints). Show that x_0 is a regular point of the constraints (1) if and only if there is no set of λ_i's, $i \in I$ satisfying

   $$\sum_{i \in I} \lambda_i g_i'(x_0) = \theta, \qquad \lambda_i \geq 0 \qquad \text{for all } i \in I, \qquad \sum_{i \in I} \lambda_i > 0.$$

5. Show that if in the generalized Kuhn-Tucker theorem X is normed, f and G are Fréchet differentiable, and the vector x is required to lie in a given convex set $\Omega \subset X$ (as well as to satisfy $G(x) \leq \theta$), then there is a $z_0^* \geq \theta$ such that $\langle G(x_0), z_0^* \rangle = 0$ and $f'(x_0) + z_0^* G'(x_0) \in [\Omega - x_0]^{\oplus}$.
6. A bomber pilot at a certain initial position above the ground seeks the path of shortest distance to put him over a certain target. Considering only two dimensions (vertical and horizontal), what is the nature of the solution to his problem when there are mountain ranges between him and his target?

7. Let X be a normed linear space and let Z be a normed linear space having positive cone P. Let G be a Fréchet differentiable mapping of X into Z. A point x_0 is said to satisfy the *Kuhn-Tucker constraint qualification* relative to the inequality $G(x) \leq \theta$ if $G(x_0) \leq \theta$ and if for every $h \in X$ satisfying $G(x_0) + G'(x_0)h \leq \theta$ there is a differentiable arc $x(t)$ defined for $t \in [0, 1]$ such that

(i) $G(x(t)) \leq \theta$ for all $t \in [0, 1]$

(ii) $\left.\frac{dx(t)}{dt}\right|_{t=0} = h$ (iii) $x(0) = x_0$.

Give an example of a finite-dimensional mapping G and a point x_0 that satisfies the Kuhn-Tucker constraint qualification but is not regular. Give an example of a finite-dimensional mapping G and an x_0, $G(x_0) \leq \theta$ that does not satisfy the Kuhn-Tucker constraint qualification.

8. Let X, Z, P, and G be as in Problem 7 and let f be a real-valued functional on X. Suppose x_0 minimizes f subject to the constraint $G(x) \leq \theta$ and that x_0 satisfies the Kuhn-Tucker constraint qualification. Show that $f'(x_0)h \geq 0$ for every h satisfying $G(x_0) + G'(x_0)h \leq \theta$. Using this result, prove a Lagrange multiplier theorem for finite-dimensional spaces.

9. Let T be a transformation mapping a vector space X into a normed space Z with positive cone P. We say that T has a *convex Gateaux differential*

$$\delta^+ T(x; h) = \lim_{\alpha \to 0+} \frac{1}{\alpha} [T(x + \alpha h) - T(x)]$$

if the limit on the right exists for all $h \in X$ and if $\delta^+ T(x; h)$ is convex in the variable h. Let f be a functional on X and G a transformation from X into Z. Assume that both f and G possess convex Gateaux differentials. Let x_0 be a solution to the problem:

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } G(x) \leq \theta. \end{aligned}$$

Assume that there exists an h such that $G(x_0) + \delta^+ G(x_0, h) < \theta$. Show that there is a $z_0^* \geq \theta$ such that

$$\delta^+ f(x_0, h) + \langle \delta^+ G(x_0, h), z_0^* \rangle \geq 0$$

for all $h \in X$. Give an example of a functional having a convex Gateaux differential but not a linear Gateaux differential.

10. After a heavy military campaign a certain army requires many new shoes. The quartermaster can order three sizes of shoes. Although he does not know precisely how many of each size are required, he feels that the demands for the three sizes are independent and the demand for each size is uniformly distributed between zero and three thousand pairs. He wishes to allocate his shoe budget of four thousand dollars among the three sizes so as to maximize the expected number of men properly shod. Small shoes cost one dollar per pair, medium shoes cost two dollars per pair, and large shoes cost four dollars per pair. How many pairs of each size should he order?
11. Because of an increasing average demand for its product, a firm is considering a program of expansion. Denoting the firm's capacity at time t by $c(t)$ and the rate of demand by $d(t)$, the firm seeks the non-decreasing function $c(t)$ starting from $c(0)$ that maximizes

$$\int_0^T \{\min [c(t), d(t)] - [\dot{c}(t)]^2\}\, dt$$

where the first term in the integrand represents revenue due to sales and the second represents expansion costs. Show that this problem can be stated as a convex programming problem. Apply the considerations of Problem 9 to this problem.
12. Derive the necessary conditions for the problem of extremizing

$$\int_{t_0}^{t_1} f(x, \dot{x}, t)\, dt$$

subject to

$$\phi(x, \dot{x}, t) \leq 0,$$

making assumptions similar to those in Example 3, Section 9.4.
13. Consider these two optimal control problems:

$$A\begin{cases}\text{minimize } \displaystyle\int_{t_0}^{t_1} l(x, u, t)\, dt \\ \text{subject to } \dot{x}(t) = f(x, u, t), \qquad x(t_0) \text{ fixed}, \qquad \displaystyle\int_{t_0}^{t_1} K(x, u, t)\, dt = b\end{cases}$$

$$B\begin{cases}\text{minimize } \psi(x(t_1)) \\ \text{subject to } \dot{x}(t) = f(x, u, t), \qquad x(t_0) \text{ fixed}, \qquad G(x(t_1)) = c.\end{cases}$$

Show that by the introduction of additional components in the state vector, a problem of type A can be converted to one of type B. Show that if G and ψ have continuous partial derivatives, a problem of type B can be converted to one of type A.

14. A discrete-time system is governed by the set of difference equations

$$x(k+1) = f(x(k), u(k)),$$

where $x(k)$ is an n vector, $u(k)$ is m-vector control, and f has continuous partial derivatives. Find a set of necessary conditions for the problem of controlling the system from a given $x(0)$ so as to minimize

$$\sum_{k=0}^{N} l(x(k), u(k)),$$

where the function l has continuous partial derivatives.

15. Using the results of Problem 14, find an optimal feedback control law when

$$f(x(k), u(k)) = Ax(k) + Bu(k)$$

$$l(x(k), u(k)) = x'(k)Qx(k) + u'(k)Ru(k),$$

where Q is positive semidefinite and R is positive definite.

16. Show that in the one-dimensional optimal control problem:

$$\text{minimize } \int_0^1 l(x, u)\, dt$$

$$\text{subject to } \dot{x}(t) = -x(t) + u^2(t)$$

$$x(0) = 1$$

$$x(1) = e^{-1},$$

the constraints are not regular.

17. Show that for the general optimal control problem discussed in Section 9.5, a Lagrangian statement with an additional scalar multiplier can be made even if the system is not regular.

18. Let X and U be normed spaces and let $A[x, u] = Bx + Cu$ where B and C are bounded linear operators with range in X. Assume that the equation $A[x, u] = \theta$ defines a unique implicit solution $x(u)$. Show that for any pair (x, u) satisfying $A[x, u] = \theta$ and any $b^* \in X^*$, $c^* \in U^*$, we have $\langle x, b^* \rangle + \langle u, c^* \rangle = \langle u, c^* - C^*\lambda^* \rangle$ where $B^*\lambda^* = b^*$. Compare with Proposition 1, Section 9.6.

REFERENCES

§9.2. This form of the inverse function theorem is apparently new, but the proof is based on a technique in Luisternik and Sobolev [101, pp. 202–208]. For standard results on inverse function theorems, see Apostol [10] and Hildebrandt and Graves [72].

§9.3. The general Lagrange multiplier theorem is due to Luisternik (in Luisternik and Sobolev [101]). Another similar result is due to Goldstine [60], [61].

§9.4. Extremization problems subject to inequality constraints in finite dimension were first treated systematically by John [76] and Kuhn and Tucker [91]. There are a number of extensions of the Kuhn-Tucker result to more general spaces, but these results are a good deal weaker than the corresponding results for equalities (see Problems 7 and 8 and Russell [135]). The approach we take is similar to that employed by Balakrishnan [15] for a finite number of inequalities. For an interesting development of necessary conditions in finite-dimensional spaces, see Canon, Cullum, and Polack [28].

§9.5. For a nice discussion of necessary conditions for optimal control, see Blum [23].

§9.6. See Pontryagin, Boltyanskii, Gamkrelidze, and Mishchenko [119] or Lee and Markus [95]. For an approach to the maximum principle closely connected with mathematical programming, see Neustadt [112], [113]. For an analysis similar to the one in this section, see Rozonoer [132], [133], [134].

10
ITERATIVE METHODS OF OPTIMIZATION

10.1 Introduction

Although a number of interesting optimization problems can be completely resolved analytically, or reduced to simple finite-dimensional problems, the great majority of problems arising from large industrial, aerospace, or governmental systems must ultimately be treated by computer methods. The reason for this is not that the necessary conditions are too difficult to derive but rather that solution of the resulting nonlinear equations is usually beyond analytic tractability.

There are two basic approaches for resolving complex optimization problems by numerical techniques: (1) formulate the necessary conditions describing the optimal solution and solve these equations numerically (usually by some iterative scheme) or (2) bypass the formulation of the necessary conditions and implement a direct iterative search for the optimum. Both methods have their merits, but at present the second appears to be the most effective since progress during the iterations can be measured by monitoring the corresponding values of the objective functional. In this chapter we introduce some of the basic concepts associated with both procedures. Sections 10.2 and 10.3 discuss methods for solving nonlinear equations; the remaining sections discuss methods for minimizing objective functionals.

The relevance of the material in the previous chapters to implementation of the first approach is obvious. In the second approach, however, since the necessary conditions are abandoned, it is perhaps not clear that any benefit is derived from classical optimization theory. Nevertheless, adjoints, Lagrange multipliers, and duality nearly always enter any detailed analysis of an iterative technique. For instance, the Lagrange multipliers of a problem are often by-products of an iterative search procedure. The most important tie between the two aspects of optimization, however, is that much of the analytical machinery and geometric insight developed for the

theory of optimization underlies much of the reasoning that leads to new, effective, computational procedures.

METHODS FOR SOLVING EQUATIONS

10.2 Successive Approximation

In its general form the classical method of successive approximation applies to equations of the form $x = T(x)$. A solution x to such an equation is said to be a *fixed point* of the transformation T since T leaves x invariant. To find a fixed point by successive approximation, we begin with an initial trial vector x_1 and compute $x_2 = T(x_1)$. Continuing in this manner iteratively, we compute successive vectors $x_{n+1} = T(x_n)$. Under appropriate conditions the sequence $\{x_n\}$ converges to a solution of the original equation.

Definition. Let S be a subset of a normed space X and let T be a transformation mapping S into S. Then T is said to be a *contraction mapping* if there is an α, $0 \leq \alpha < 1$ such that $\|T(x_1) - T(x_2)\| \leq \alpha \|x_1 - x_2\|$ for all $x_1, x_2 \in S$.

Note for example that a transformation having $\|T'(x)\| \leq \alpha < 1$ on a convex set S is a contraction mapping since, by the mean value inequality, $\|T(x_1) - T(x_2)\| \leq \sup \|T'(x)\| \|x_1 - x_2\| \leq \alpha \|x_1 - x_2\|$.

Theorem 1. (*Contraction Mapping Theorem*) *If T is a contraction mapping on a closed subset S of a Banach space, there is a unique vector $x_0 \in S$ satisfying $x_0 = T(x_0)$. Furthermore, x_0 can be obtained by the method of successive approximation starting from an arbitrary initial vector in S.*

Proof. Select an arbitrary element $x_1 \in S$. Define the sequence $\{x_n\}$ by the formula $x_{n+1} = T(x_n)$. Then $\|x_{n+1} - x_n\| = \|T(x_n) - T(x_{n-1})\| \leq \alpha \|x_n - x_{n-1}\|$. Therefore,

$$\|x_{n+1} - x_n\| \leq \alpha^{n-1} \|x_2 - x_1\|.$$

It follows that

$$\begin{aligned} \|x_{n+p} - x_n\| &\leq \|x_{n+p} - x_{n+p-1}\| + \|x_{n+p-1} - x_{n+p-2}\| + \cdots + \|x_{n+1} - x_n\| \\ &\leq (\alpha^{n+p-2} + \alpha^{n+p-3} + \cdots + \alpha^{n-1}) \|x_2 - x_1\| \\ &\leq (\alpha^{n-1} \sum_{k=0}^{\infty} \alpha^k) \|x_2 - x_1\| = \frac{\alpha^{n-1}}{1 - \alpha} \|x_2 - x_1\|, \end{aligned}$$

and hence we conclude that $\{x_n\}$ is a Cauchy sequence. Since S is a closed subset of a complete space, there is an element $x_0 \in S$ such that $x_n \to x_0$.

We now show that $x_0 = T(x_0)$. We have

$$\|x_0 - T(x_0)\| = \|x_0 - x_n + x_n - T(x_0)\| \leq \|x_0 - x_n\| + \|x_n - T(x_0)\|$$
$$\leq \|x_0 - x_n\| + \alpha\|x_{n-1} - x_0\|.$$

By appropriate choice of n the right-hand side of the above inequality can be made arbitrarily small. Thus $\|x_0 - T(x_0)\| = 0$; $x_0 = T(x_0)$.

It remains only to show that x_0 is unique. Assume that x_0 and y_0 are fixed points. Then

$$\|x_0 - y_0\| = \|T(x_0) - T(y_0)\| \leq \alpha\|x_0 - y_0\|.$$

Thus $x_0 = y_0$. ∎

The process of successive approximation is often illustrated by a diagram such as that of Figure 10.1. The figure represents the process of solving the one-dimensional equation $x = f(x)$. On the diagram this is equivalent to finding the point of intersection of $f(x)$ with the forty-five degree line through the origin. Starting with x_1, we derive $x_2 = f(x_1)$ by moving along the curve as shown. The f shown in the figure has slope less than unity and is thus a contraction. Figure 10.2 shows a case where successive approximation diverges.

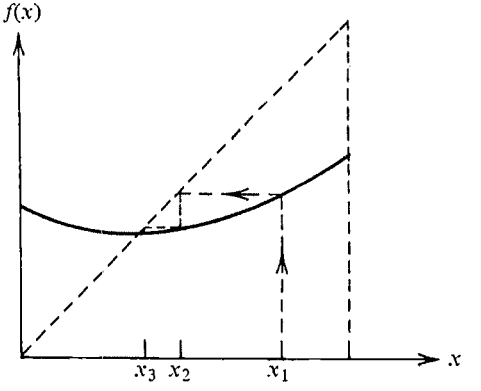

Figure 10.1 Successive approximation

Example 1. (Linear Algebraic Equations) Consider the set of equations $Ax = b$ where A is an $n \times n$ matrix. A sequence of approximate solutions $x_k = (x_1^k, x_2^k, \ldots, x_n^k), k = 1, 2, \ldots$, can be generated by solving the equations

$$\begin{aligned}
a_{11}x_1^{k+1} + a_{12}x_2^k &+ \cdots + a_{1n}x_n^k &&= b_1 \\
a_{21}x_1^k + a_{22}x_2^{k+1} &+ \cdots + a_{2n}x_n^k &&= b_2 \\
&\vdots &&\vdots \\
a_{n1}x_1^k &+ \cdots + a_{nn}x_n^{k+1} &&= b_n
\end{aligned}$$

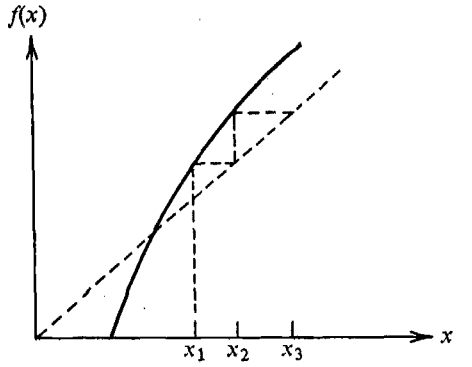

Figure 10.2 Divergent case of successive approximation

for x_{k+1} given x_k. In other words, the p-th equation is solved for the new component x_p^{k+1} by first setting all other components equal to their values at the last iteration. We analyze this method and show that it converges if A has a property called strict diagonal dominance.

Definition. A matrix A is said to have *strict diagonal dominance* if

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|$$

for each i.

In what follows we assume that A has strict diagonal dominance and that each of the n equations represented by $Ax = b$ has been appropriately scaled so that $a_{ii} = 1$ for each i. The equation may be rewritten as

$$x = (I - A)x + b.$$

Defining $(I - A)x + b = T(x)$, the problem is equivalent to that of finding a fixed point of T. Furthermore, the method of successive approximation proposed above for this problem is equivalent to ordinary successive approximation applied to T. Thus it is sufficient to show that T is a contraction mapping with respect to some norm on n-dimensional space.

Let X be the space of n-tuples with norm defined by

$$\|x\| = \max_{1 \leq i \leq n} |x_i|.$$

This norm on X induces a norm on $n \times n$ matrices B:

$$\|B\| = \max_i \sum_{j=1}^{n} |b_{ij}|.$$

For the mapping T defined above we have

$$\|T(x) - T(y)\| = \|(A - I)(x - y)\| \leq \|A - I\| \|x - y\|.$$

However, since $a_{ii} = 1$, the norm of $A - I$ is

$$\|A - I\| = \max_i \sum_{j \neq i} |a_{ij}| = \alpha.$$

By the assumption of strict diagonal dominance, $\alpha < 1$ and thus T is a contraction mapping.

Example 2. Consider the integral equation

$$x(t) = f(t) + \lambda \int_a^b K(t, s)x(s)\, ds.$$

Let $\int_a^b \int_a^b K^2(s, t)\, dt\, ds = \beta^2 < \infty$, and assume that $f \in X = L_2[a, b]$. Then the integral on the right-hand side of the integral equation defines a bounded linear operator on X having norm less than or equal to β. It follows that the mapping

$$T(x) = f(t) + \lambda \int_a^b K(t, s)x(s)\, ds$$

is a contraction mapping on X provided that $|\lambda| < 1/\beta$. Thus, for this range of the parameter λ the equation has a unique solution which can be determined by successive approximation.

The basic idea of successive approximation and contraction mappings can be modified in several ways to produce convergence theorems for a number of different situations. We consider one such modification below. Others can be found in the problems at the end of this chapter.

Theorem 2. *Let T be a continuous mapping from a closed subset S of a Banach space into S, and suppose that T^n is a contraction mapping for some positive integer n. Then T has a unique fixed point in S which can be found by successive approximation.*

Proof. Let x_1 be arbitrary in S. Define the sequence $\{x_i\}$ by

$$x_{i+1} = T(x_i).$$

Now since T^n is a contraction, it follows by Theorem 1 that the subsequence $\{x_{nk}\}$ converges to an element $x_0 \in S$ which is a fixed point of T^n. We show that x_0 is a unique fixed point of T.

By the continuity of T, the element $T(x_0)$ can be obtained by applying T^n successively to $T(x_1)$. Therefore, we have $x_0 = \lim_{k \to \infty} T^{nk}(x_1)$ and

$T(x_0) = T[\lim_{k\to\infty} T^{nk}(x_1)] = \lim_{k\to\infty} T^{nk}[T(x_1)]$. Hence, again using the continuity of T,

$$\begin{aligned}\|x_0 - T(x_0)\| &= \lim_{k\to\infty} \|T^{nk}(x_1) - T^{nk}[T(x_1)]\| \\ &= \lim_{k\to\infty} \|T^n\{T^{n(k-1)}(x_1) - T^{n(k-1)}[T(x_1)]\}\| \\ &\leq \alpha \lim_{k\to\infty} \|T^{n(k-1)}(x_1) - T^{n(k-1)}[T(x_1)]\| \\ &= \alpha\|x_0 - T(x_0)\|,\end{aligned}$$

where $\alpha < 1$. Thus $x_0 = T(x_0)$.

If x_0, y_0 are fixed points, then $\|x_0 - y_0\| = \|T^n(x_0) - T^n(y_0)\| \leq \alpha\|x_0 - y_0\|$ and hence $x_0 = y_0$. Thus the x_0 found by successive approximation is a unique fixed point of T. ▮

Example 3. (Ordinary Differential Equations) Consider the ordinary differential equation

$$\dot{x}(t) = f[x(t), t].$$

The function x may be taken to be scalar valued or vector valued, but for simplicity we assume here that it is scalar valued. Suppose that $x(t_0)$ is specified. We seek a solution $x(t)$ for $t_0 \leq t \leq t_1$.

We show that under the assumption that the function f satisfies a Lipschitz condition on $[t_0, t_1]$ of the form

$$|f[x_1, t] - f[x_2, t]| \leq M|x_1 - x_2|,$$

a unique solution to the initial value problem exists and can be found by successive approximation.

The differential equation is equivalent to the integral equation

$$x(t) = x_0 + \int_{t_0}^{t} f[x(\tau), \tau]\, d\tau.$$

On the space $X = C[t_0, t_1]$ let the mapping T be defined as

$$T(x) = \int_{t_0}^{t} f[x(\tau), \tau]\, d\tau.$$

Then

$$\begin{aligned}\|T(x_1) - T(x_2)\| &= \left\| \int_{t_0}^{t} \{f[x_1, \tau] - f[x_2, \tau]\}\, d\tau \right\| \\ &\leq \left\| \int_{t_0}^{t} M\|x_1 - x_2\|\, d\tau \right\| \leq M(t_1 - t_0)\|x_1 - x_2\|.\end{aligned}$$

Thus T is a contraction mapping if $M < 1/(t_1 - t_0)$. A simple calculation, however, shows that

$$\|T^n(x_1) - T^n(x_2)\| \leq \frac{M^n(t_1 - t_0)^n}{n!} \|x_1 - x_2\|.$$

Since $n!$ increases faster than any geometric progression, it follows that for sufficiently large n, T^n is a contraction mapping. Therefore, Theorem 2 applies and there is a unique solution to the differential equation which can be obtained by successive approximation.

A slight modification of this technique can be used to solve the Euler-Lagrange differential equation arising from certain optimal control problems. See Problem 5.

For a successive approximation procedure applied to a contraction mapping T having fixed point x_0, we have the inequalities

$$\|x_{n+1} - x_n\| \leq \alpha \|x_n - x_{n-1}\|, \qquad \alpha < 1, \tag{1}$$

and

$$\|x_n - x_0\| \leq \alpha \|x_{n-1} - x_0\|. \tag{2}$$

A sequence $\{x_n\}$ is said to converge *linearly* to x_0 if

$$\limsup \frac{\|x_n - x_0\|}{\|x_{n-1} - x_0\|} = \alpha$$

for some α, $0 < \alpha < 1$. Thus, in particular, (2) implies that a successive approximation procedure converges linearly. In many applications, however, linear convergence is not sufficiently rapid so faster techniques must be considered.

10.3 Newton's Method

Newton's method is an iterative technique for solving an equation of the form $P(x) = \theta$. As originally conceived, it applies to equations of a single real variable but it has a direct extension applicable to nonlinear transformations on normed spaces.

The basic technique for a function of a real variable is illustrated in Figure 10.3. At a given point the graph of the function P is approximated by its tangent, and an approximate solution to the equation $P(x) = 0$ is taken to be the point where the tangent crosses the x axis. The process is

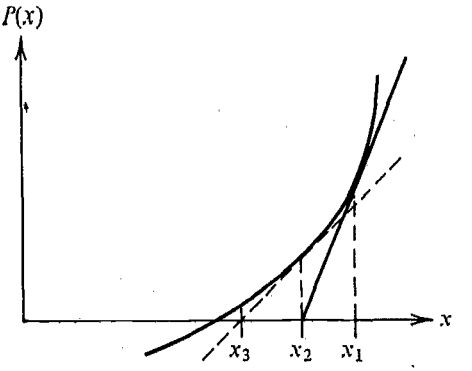

Figure 10.3 Newton's method

then repeated from this new point. This procedure defines a sequence of points according to the recurrence relation

$$x_{n+1} = x_n - \frac{P(x_n)}{P'(x_n)}.$$

Example 1. Newton's method can be used to develop an effective iterative scheme for computing square roots. Letting $P(x) = x^2 - a$, we obtain by Newton's method

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \tfrac{1}{2}\left[x_n + \frac{a}{x_n}\right].$$

This algorithm converges quite rapidly, as illustrated below, for the computation of $\sqrt{10}$, beginning with the initial approximation $x_1 = 3.0$.

Iteration	x_n	x_n^2
1	3.0000000000	9.0000000000
2	3.1666666667	10.0277777778
3	3.1622807018	10.0000192367
4	3.1622776602	10.0000000000

When applied to equations of form $P(x) = \theta$, where P is a nonlinear operator between Banach spaces, Newton's method becomes

$$x_{n+1} = x_n - [P'(x_n)]^{-1}P(x_n).$$

An interpretation of the method is, of course, that the original equation is linearized about the point x_n and then solved for x_{n+1}. Alternatively, the method can be viewed as the method of successive approximation applied

to the operator $T(x) \equiv x - [P'(x)]^{-1}P(x)$. A fixed point of T is a solution of $P(x) = \theta$.

When analyzing the convergence of Newton's method, we assume that P is twice Fréchet differentiable throughout the region of interest. Corresponding to a point x_n, we denote by p_n the bounded linear operator $P'(x_n)$ and by p_n^{-1} its inverse $[P'(x_n)]^{-1}$ if it exists. Since Newton's method amounts to successive approximation with $T(x) = x - [P'(x)]^{-1}P(x)$, an initial approach at an analysis of convergence would be to determine if $\|T'(x)\| < 1$. Since

$$T'(x_n) = p_n^{-1}P''(x_n)p_n^{-1}[P(x_n)],$$

if $\|p_n^{-1}\| \leq \beta$, $\|P''(x_n)\| \leq K$, $\|p_n^{-1}[P(x_n)]\| \leq \eta$, $h = \beta\eta K$, we have $\|T'(x_n)\| \leq h$. Therefore, by the contraction mapping principle, we expect to obtain convergence if $h < 1$ for every point in the region of interest. In the following theorem it is shown that if $h < \frac{1}{2}$ at the initial point, then $h < \frac{1}{2}$ holds for all points in the iteration and Newton's method converges.

Just as with the principle of contraction mapping, study of the convergence of Newton's method answers some questions concerning existence and uniqueness of a solution to the original equation.

Theorem 1. *Let X and Y be Banach spaces and let P be a mapping from X to Y. Assume further that:*

1. *P is twice Fréchet differentiable and that $\|P''(x)\| \leq K$.*
2. *There is a point $x_1 \in X$ such that $p_1 = P'(x_1)$ has a bounded inverse p_1^{-1} with $\|p_1^{-1}\| \leq \beta_1$, $\|p_1^{-1}[P(x_1)]\| \leq \eta_1$.*
3. *The constant $h_1 = \beta_1\eta_1 K$ satisfies $h_1 < \frac{1}{2}$.*

Then the sequence $x_{n+1} = x_n - p_n^{-1}[P(x_n)]$ exists for all $n > 1$ and converges to a solution of $P(x) = \theta$.

Proof. We show that if the point x_1 satisfies 1, 2, and 3, the point $x_2 = x_1 - p_1^{-1}P(x_1)$ satisfies the same conditions with new constants β_2, η_2, h_2.

Clearly, x_2 is well defined and $\|x_2 - x_1\| \leq \eta_1$. We have, by the mean value inequality,

$$\|p_1^{-1}[p_1 - p_2]\| \leq \beta_1 \sup_{\bar{x}} \|P''(\bar{x})\| \|x_2 - x_1\|,$$

where $\bar{x} = x_1 + \alpha(x_2 - x_1)$, $0 \leq \alpha \leq 1$. Thus $\|p_1^{-1}[p_1 - p_2]\| \leq \beta_1 K\eta_1 = h_1$. Since $h_1 < \frac{1}{2}$, it follows that the linear operator

$$H = I - p_1^{-1}[p_1 - p_2] = p_1^{-1}p_2$$

has a bounded inverse satisfying $\|H^{-1}\| \leq 1/(1-h_1)$ (see Problem 4). We have $p_1 H = p_2$ and $(p_1 H)^{-1} = H^{-1} p_1^{-1}$ so p_2^{-1} exists. An estimate of its bound is

$$\|p_2^{-1}\| \leq \|H^{-1}\| \|p_1^{-1}\| \leq \frac{\beta_1}{1-h_1} = \beta_2.$$

To obtain a bound for $\|p_2^{-1} P(x_2)\|$, we consider the operator $T_1(x) = x - p_1^{-1} P(x)$. Clearly, $T_1(x_1) = x_2$ and $T_1'(x_1) = \theta$. Thus

$$p_1^{-1} P(x_2) = T_1(x_1) - T_1(x_2) - T_1'(x_1)(x_2 - x_1).$$

By applying Proposition 3, Section 7.3, we obtain

$$\begin{aligned} \|p_1^{-1} P(x_2)\| &\leq \tfrac{1}{2} \sup \|T''(x)\| \|x_2 - x_1\|^2 \\ &= \tfrac{1}{2} \sup \|p_1^{-1} P''(x)\| \|x_2 - x_1\|^2 \\ &\leq \tfrac{1}{2} \beta_1 K \eta_1^2 = \tfrac{1}{2} h_1 \eta_1. \end{aligned}$$

Therefore,

$$\|p_2^{-1} P(x_2)\| = \|H^{-1} p_1^{-1} P(x_2)\| < \tfrac{1}{2} \frac{h_1 \eta_1}{1-h_1} = \eta_2 < \tfrac{1}{2} \eta_1.$$

Finally, setting $h_2 = \beta_2 \eta_2 K$, we have

$$h_2 \leq \tfrac{1}{2} \frac{h_1^2}{(1-h_1)^2} < \tfrac{1}{2}.$$

Hence the conditions 1, 2, and 3 are satisfied by the point x_2 and the constants β_2, η_2, and h_2. It follows by induction that Newton's process defines $\{x_n\}$.

Since $\eta_{n+1} < \frac{1}{2}\eta_n$, it follows that $\eta_n < (\frac{1}{2})^{n-1} \eta_1$. Also since $\|x_{n+1} - x_n\| < \eta_n$ it follows that $\|x_{n+k} - x_n\| < 2\eta_n$ and hence that the sequence $\{x_n\}$ converges to a point $x_0 \in X$.

To prove that x_0 satisfies $P(x_0) = \theta$, we note that the sequence $\{\|p_n\|\}$ is bounded since

$$\|p_n\| \leq \|p_1\| + \|p_n - p_1\| \leq \|p_1\| + K\|x_n - x_1\|$$

and the sequence $\{\|x_n - x_1\|\}$ is bounded since it is convergent. Now for each n, $p_n(x_{n+1} - x_n) + P(x_n) = \theta$; and since $\|x_{n+1} - x_n\| \to 0$ and $\|p_n\|$ is bounded, it follows that $\|P(x_n)\| \to 0$. By the continuity of P, $P(x_0) = \theta$. ∎

It is assumed in the above theorem that P is defined throughout the Banach space X and that $\|P''(x)\| \leq K$ everywhere. It is clear that these requirements are more severe than necessary since they are only used in the neighborhood of the points of the successive approximations. It can be

shown that if $P(x)$ is defined and $\|P''(x)\| \leq K$ in a neighborhood of x_1 with radius

$$r > \frac{1}{h_1}(1 - \sqrt{1 - 2h_1})\eta_1,$$

the successive approximations of Newton's method converge and remain within this neighborhood.

The above theorem can be paraphrased roughly by simply saying that Newton's method converges provided that the initial approximation x_1 is sufficiently close to the solution x_0. Because if $[P'(x)]^{-1}$ is bounded near x_0, the quantity $\eta = [P'(x)]^{-1}P(x)$ goes to zero as $x \to x_0$; therefore $h = \beta\eta K$ is small for x close to x_0.

The most attractive feature of Newton's method, the reward for the high price paid for the possibly difficult job of solving a linearized version at each step, is its rate of convergence. Suppose that Newton's method converges to a solution $x_0 \in X$ where $[P'(x_0)]^{-1}$ exists. Furthermore, assume that, within an open region R containing x_0 and the sequence $\{x_n\}$, the quantities $\|[P'(x)]^{-1}\|$, $\|P''(x)\|$, and $\|P'''(x)\|$ are bounded. Then again defining $T(x) = x - [P'(x)]^{-1}P(x)$, we have

$$\begin{aligned} x_{n+1} - x_0 &= x_n - [P'(x_n)]^{-1}P(x_n) - x_0 \\ &= x_n - [P'(x_n)]^{-1}P(x_n) - \{x_0 - [P'(x_0)]^{-1}P(x_0)\} \\ &= T(x_n) - T(x_0). \end{aligned}$$

Since $T'(x_0) = \theta$,

$$\|x_{n+1} - x_0\| \leq \tfrac{1}{2} \sup_{\bar{x}} \|T''(\bar{x})\| \|x_n - x_0\|^2,$$

where $\bar{x} = x_n + \alpha(x_n - x_0)$, $0 \leq \alpha \leq 1$. Hence,

$$\|x_{n+1} - x_0\| \leq c\|x_n - x_0\|^2, \tag{1}$$

where $c = \frac{1}{2} \sup_{x \in R} \|T''(x)\|$, a bound depending upon $\|P'''(x)\|$. Relation (1) is referred to as *quadratic convergence.*

The overall conclusion of the analysis of Newton's method is that, under mild restrictions, the method converges quadratically provided that the initial approximation is sufficiently near the solution. In practice the detailed criteria for convergence stated in this section are difficult to check, and it is often simpler to carry out the process than to verify beforehand that it will converge. Moreover, the method may converge even though the sufficiency conditions are violated. One device useful in these situations is to begin iterating with a slower but surer technique and then change over to Newton's method to gain the advantage of quadratic convergence near the end of the process.

Example 2. (Two-Point Boundary Value Problem) Newton's method can be used very effectively to compute solutions of nonlinear two-point boundary value problems such as those arising in connection with optimal control problems.

Consider first the linear two-point boundary value problem

$$\dot{x}(t) = A(t)x(t) + v(t), \tag{2}$$

where $x(t)$ is an n-dimensional vector function of t subject to the boundary conditions

$$Cx(t_1) = c_1 \tag{3}$$

$$Dx(t_2) = d_2, \tag{4}$$

where $\dim(c_1) + \dim(d_2) = n$. Since the system is linear, we may write the superposition relation

$$x(t_2) = \Phi(t_2, t_1)x(t_1) + \int_{t_1}^{t_2} \Phi(t_2, t)v(t)\,dt, \tag{5}$$

where $\Phi(t_2, t)$ is the matrix function of t satisfying

$$\dot{\Phi}(t_2, t) = -\Phi(t_2, t)A(t)$$

$$\Phi(t_2, t_2) = I.$$

Note that $\Phi(t_2, t)$ can be found by integrating backward from t_2.

Defining $b = \int_{t_1}^{t_2} \Phi(t_2, t)v(t)\,dt$, the boundary conditions (3) and (4) can be expressed entirely in terms of t_1

$$Cx(t_1) = c_1$$

$$D\Phi(t_2, t_1)x(t_1) = d_2 - Db.$$

Therefore, assuming the existence of the appropriate inverse, we have

$$x(t_1) = \begin{bmatrix} C \\ D\Phi(t_2, t_1) \end{bmatrix}^{-1} \begin{matrix} c_1 \\ d_2 - Db \end{matrix}\Bigg]. \tag{6}$$

Having determined $x(t_1)$, the original equation (2) can be solved by a single forward integration.

Now consider a similar, nonlinear, two-point boundary problem:

$$\dot{x}(t) = F(x, t)$$

$$Cx(t_1) = c_1$$

$$Dx(t_2) = d_2.$$

Although this problem cannot be solved by a single integration or by superposition, it can often be solved iteratively by Newton's method. We start with an initial approximation $x_1(t)$ and define

$$\dot{x}_{n+1}(t) = F(x_n, t) + F_x(x_n, t)(x_{n+1}(t) - x_n(t))$$
$$Cx_{n+1}(t_1) = c_1$$
$$Dx_{n+1}(t_2) = d_2 .$$

At each step of the iteration the linearized version of the problem is solved by the method outlined above. Then provided that the initial approximation is sufficiently close to the solution, we can expect this method to converge quadratically to the solution.

DESCENT METHODS

10.4 General Philosophy

Successive approximation, Newton's method, and other methods for solving nonlinear equations, when applied to an optimization problem, iterate on the equations derived as necessary conditions for an optimal solution. A major disadvantage of this approach is that these iterative techniques may converge only if the initial approximation is sufficiently close to the solution. With these methods only local convergence is guaranteed.

A more direct approach for optimization problems is to iterate in such a way as to decrease the cost functional continuously from one step to the next. In this way global convergence, convergence from an arbitrary starting point, often can be insured.

As a general framework for the method, assume that we seek to minimize a functional f and that an initial point x_1 is given. The iterations are constructed according to an equation of the form

$$x_{n+1} = x_n + \alpha_n p_n ,$$

where α_n is a scalar and p_n is a (direction) vector. The procedure for selecting the vector p_n varies from technique to technique but, ideally, once it is chosen the scalar α_n is selected to minimize $f(x_n + \alpha p_n)$, regarded as a function of the scalar α. Generally, things are arranged (by multiplying p_n by -1 if necessary) so that $f(x_n + \alpha p_n) < f(x_n)$ for small positive α. The scalar α_n is then often taken as the smallest positive root of the equation

$$\frac{d}{d\alpha} f(x_n + \alpha p_n) = 0.$$

In practice, of course, it is rarely possible to evaluate the minimizing α exactly. Instead, some iterative search or approximation is required. The essential point, however, is that after an α_n is selected, $f(x_n + \alpha_n p_n)$ is evaluated to verify that the objective has in fact decreased from $f(x_n)$. If f has not decreased, a new value of α_n is chosen.

The descent process can be visualized in the space X where the functional f is represented by its contours. Starting from a point x_1, one moves along the direction vector p_1 until reaching, as illustrated in Figure 10.4, the first point where the line $x_1 + \alpha p_1$ is tangent to a contour of f. Alternatively, the method can be visualized, as illustrated in Figure 10.5, in the space $R \times X$, the space containing the graph of f.

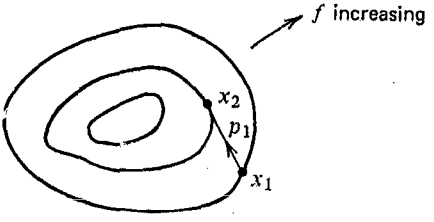

Figure 10.4 The descent process in X

If f is bounded below, it is clear that the descent process defines a bounded decreasing sequence of functional values and hence that the objective values tend toward a limit f_0. The difficulties remaining are those of insuring that f_0 is, in fact, the minimum of f, that the sequence of approximations $\{x_n\}$ converges to a minimizing vector, and finally, the most difficult, that convergence is rapid enough to make the whole scheme practical.

Example 1. Newton's method can be modified for optimization problems to become a rapidly converging descent method. Suppose again that we seek to minimize the functional f on a Banach space X. This might be accomplished by the ordinary Newton's method for solving the nonlinear equation $F(x) = \theta$ where $F(x) = f'(x)$, but this method suffers from the lack of a global convergence theorem. The method is modified to become the Newtonian descent method by selecting the direction vectors according to the ordinary Newton's method but moving along them to a point minimizing f in that direction. Thus the general iteration formula is

$$x_{n+1} = x_n - \alpha_n[F'(x_n)]^{-1}F(x_n) = x_n - \alpha_n[f''(x_n)]^{-1}f'(x_n)$$

and α_n is chosen to minimize $f(x_{n+1})$.

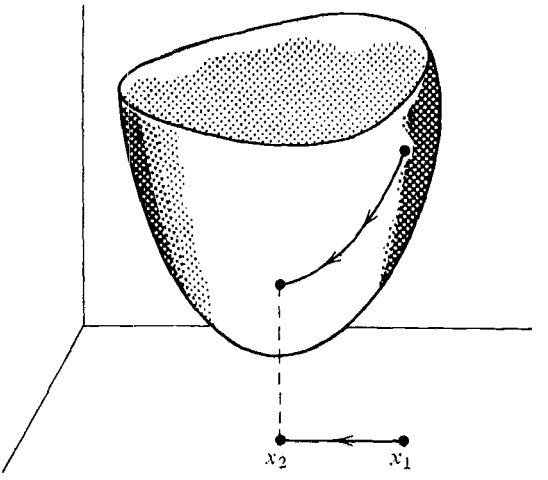

Figure 10.5 The descent process in $R \times X$

10.5 Steepest Descent

The most widely used descent procedure for minimizing a functional f, the method of steepest descent, is applicable to functionals defined on a Hilbert space X. In this method the direction vector p_n at a given point x_n is chosen to be the negative of the gradient of f at x_n. If X is not a Hilbert space, the method can be modified by selecting p_n to be a vector aligned with, or almost aligned with, the negative gradient. In this section, however, we restrict our attention to functionals on a Hilbert space.

An application of the method is to the minimization of a quadratic functional

$$f(x) = (x \mid Qx) - 2(b \mid x),$$

where Q is a self-adjoint positive-definite operator on the Hilbert space X. This problem is of particular theoretical interest because it is the only problem for which a detailed convergence analysis of steepest descent and other iterative methods is available. The problem therefore provides a comparison point for the several methods. Of course the quadratic problem is of practical interest as well, as illustrated in Chapters 3 and 4.

In analyzing the quadratic problem it is assumed that the constants

$$m = \inf_{x \neq \theta} \frac{(x \mid Qx)}{(x \mid x)}$$

$$M = \sup_{x \neq \theta} \frac{(x \mid Qx)}{(x \mid x)}$$

are positive, finite numbers. Under these conditions f is minimized by the unique vector x_0 satisfying the equation

$$Qx_0 = b \tag{1}$$

and, indeed, minimization of f is completely equivalent to solving the linear equation (1). It is convenient therefore to regard any approximation x to the point minimizing f as an approximate solution to equation (1). The vector

$$r = b - Qx$$

is called the residual of the approximation; inspection of f reveals that $2r$ is the negative gradient of f at the point x.

The method of steepest descent applied to f therefore takes the form

$$x_{n+1} = x_n + \alpha_n r_n,$$

where $r_n = b - Qx_n$ and α_n is chosen to minimize $f(x_{n+1})$. The value of α_n can be found explicitly since

$$\begin{aligned} f(x_{n+1}) &= (x_n + \alpha r_n \mid Q(x_n + \alpha r_n)) - 2(x_n + \alpha r_n \mid b) \\ &= \alpha^2(r_n \mid Qr_n) - 2\alpha(r_n \mid r_n) + (x_n \mid Qx_n) - 2(x_n \mid b), \end{aligned}$$

which is minimized by

$$\alpha_n = \frac{(r_n \mid r_n)}{(r_n \mid Qr_n)}. \tag{2}$$

Steepest descent for $f(x) = (x \mid Qx) - 2(x \mid b)$ therefore progresses according to

$$x_{n+1} = x_n + \frac{(r_n \mid r_n)}{(r_n \mid Qr_n)} r_n, \tag{3}$$

where $r_n = b - Qx_n$.

Theorem 1. *For any $x_1 \in X$ the sequence $\{x_n\}$ defined by* (3) *converges (in norm) to the unique solution x_0 of $Qx = b$. Furthermore, defining*

$$F(x) = (x - x_0 \mid Q(x - x_0))$$

the rate of convergence satisfies

$$(y_n \mid y_n) \leq \frac{1}{m} F(x_n) \leq \frac{1}{m} \left(1 - \frac{m}{M}\right)^{n-1} F(x_1)$$

where $y_n = x_0 - x_n$.

Proof. Note that

$$F(x) = (x - x_0 \mid Q(x - x_0)) = (x \mid Qx) - 2(x \mid b) + (x_0 \mid Qx_0)$$
$$= f(x) + (x_0 \mid Qx_0)$$

so that both f and F achieve a minimum at x_0 and the gradients of f and F are equal.

We have

$$\frac{F(x_n) - F(x_{n+1})}{F(x_n)} = \frac{2\alpha_n(r_n \mid Qy_n) - \alpha_n^2(r_n \mid Qr_n)}{(y_n \mid Qy_n)}$$

by direct calculation. Now $r_n = Qy_n$, so in terms of r_n

$$\frac{F(x_n) - F(x_{n+1})}{F(x_n)} = \frac{\dfrac{2(r_n \mid r_n)^2}{(r_n \mid Qr_n)} - \dfrac{(r_n \mid r_n)^2}{(r_n \mid Qr_n)}}{(Q^{-1}r_n \mid r_n)}$$
$$= \frac{(r_n \mid r_n)}{(r_n \mid Qr_n)} \cdot \frac{(r_n \mid r_n)}{(Q^{-1}r_n \mid r_n)}.$$

Using $(r_n \mid Qr_n) \leq M(r_n \mid r_n)$ and $(Q^{-1}r_n \mid r_n) \leq \dfrac{1}{m}(r_n \mid r_n)$, which follows from the definition of m (see Problem 10), we obtain

$$\frac{F(x_n) - F(x_{n+1})}{F(x_n)} \geq \frac{m}{M}$$

$$\frac{F(x_{n+1})}{F(x_n)} \leq 1 - \frac{m}{M}$$

$$F(x_n) \leq \left(1 - \frac{m}{M}\right)^{n-1} F(x_1).$$

And finally,

$$(y_n \mid y_n) \leq \frac{1}{m} F(x_n) \leq \frac{1}{m}\left(1 - \frac{m}{M}\right)^{n-1} F(x_1). \quad \blacksquare$$

This process of steepest descent is illustrated for a two-dimensional problem in Figure 10.6. Note that, according to Theorem 1 and from the figure, the rate of convergence depends on the eccentricity of the elliptical contours of f. For $m = M$ the contours are circular and convergence occurs in one step.

Example 1. Consider the problem of solving the set of linear equations $Ax = b$ where A is an $N \times N$ positive-definite matrix. We assume that the

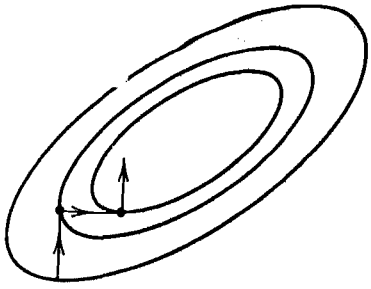

Figure 10.6 Steepest descent

equations have been scaled so that $a_{ii} = 1$, $i = 1, 2, \ldots, N$. According to the method of steepest descent (with $Q = A$), the approximate solutions are generated by

$$x_{n+1} = x_n + \alpha_n r_n.$$

Suppose that for simplicity α_n, instead of being calculated according to equation (2), is taken as $\alpha_n \equiv 1$. Then the method becomes

$$x_{n+1} = x_n + r_n = x_n + b - Ax_n,$$

or $x_{n+1} = (I - A)x_n + b$. This last equation is equivalent to the method of successive approximation given in Example 1, Section 10.2.

The method of steepest descent is frequently applied to nonquadratic problems with great success; indeed, the method rarely fails to converge to at least a local minimum. The following theorem establishes conditions for which success is guaranteed.

Theorem 2. *Let f be a functional bounded below and twice Fréchet differentiable on a Hilbert space H. Given $x_1 \in H$, let S be the closed convex hull of $\{x : f(x) < f(x_1)\}$. Assume that $f''(x)$ is self-adjoint and satisfies $0 < mI \leq f''(x) \leq MI$ throughout S (i.e., $f''(x)$ is uniformly bounded and uniformly positive definite). If $\{x_n\}$ is the sequence generated by steepest descent applied to f starting at x_1, then $f'(x_n) \to \theta$. Furthermore, there exists an $x_0 \in S$ such that $x_n \to x_0$ and $f(x_0) = \inf\{f(x) : x \in H\}$.*

Proof. Given $x \in S$, let us apply Taylor's expansion with remainder to the function $g(t) = f(tx + (1 - t)x_1)$ obtaining

$$g(1) - g(0) - g'(0) = \tfrac{1}{2} g''(\bar{t})$$

for some $\bar{t}$, $0 < \bar{t} < 1$. This leads immediately to

$$f(x) - f(x_1) - f'(x_1)(x - x_1) \geq \tfrac{1}{2} m \|x - x_1\|^2$$

from which it follows that S is bounded. Therefore the steepest-descent process defines a bounded sequence $\{x_n\}$.

As for any descent process, the corresponding sequence $\{f(x_n)\}$ is non-increasing and bounded below and therefore converges to some value f_0. Now assume that the sequence of gradients $\{f'(x_n)\}$ does not converge to zero. Then there is an $\varepsilon > 0$ such that for any N there is an $n > N$ with $\|f'(x_n)\| \geq \varepsilon$. Thus, choose n so large that $\|f'(x_n)\| \geq \varepsilon$ and $|f(x_n) - f_0| < \varepsilon^2/4M$. For $\alpha > 0$ let $x_\alpha = x_n - \alpha f'(x_n)$. We have by Taylor's expansion with remainder

$$f(x_\alpha) - f(x_n) \leq -\alpha \|f'(x_n)\|^2 + \frac{\alpha^2}{2} \|f''(\bar{x})\| \|f'(x_n)\|^2$$

$$\leq \left(-\alpha + \frac{\alpha^2}{2} M\right) \|f'(x_n)\|^2,$$

where $\bar{x} = tx_n + (1-t)x_\alpha$, $0 \leq t \leq 1$. Therefore, for $\alpha = 1/M$ we have

$$f(x_\alpha) - f(x_n) \leq -\frac{\varepsilon^2}{2M},$$

which implies that $f(x_{n+1}) < f_0$. Since this is impossible, it follows that $\|f'(x_n)\| \to 0$.

For any $x, y \in S$ we have, by the one-dimensional mean value theorem,

$$(f'(x) - f'(y) \mid x - y) = (x - y \mid f''(\bar{x})(x - y)) \geq m \|x - y\|^2,$$

where $\bar{x} = tx + (1-t)y$, $0 \leq t \leq 1$. Thus

$$\|x_{n+k} - x_n\|^2 \leq \frac{1}{m} (f'(x_{n+k}) - f'(x_n) \mid x_{n+k} - x_n)$$

$$\leq \frac{1}{m} \|f'(x_{n+k}) - f'(x_n)\| \|x_{n+k} - x_n\|$$

or

$$\|x_{n+k} - x_n\| \leq \frac{1}{m} \|f'(x_{n+k}) - f'(x_n)\|.$$

Since $\{f'(x_n)\}$ is a Cauchy sequence, so is $\{x_n\}$; thus there exists $x_0 \in S$ with $x_n \to x_0$.

Obviously, $f'(x_0) = \theta$. Given h such that $x_0 + h \in S$, there is t, $0 < t < 1$, such that

$$f(x_0 + h) = f(x_0) + \tfrac{1}{2}(h \mid f''(x_0 + th)h)$$

$$\geq f(x_0) + \frac{m}{2} \|h\|^2$$

so x_0 minimizes f in S and hence in H. ∎

Example 2. Steepest descent can be applied to the optimal control problem

$$\text{minimize} \quad J = \int_{t_0}^{t_1} l(x(t), u(t))\, dt$$

$$\text{subject to } \dot{x}(t) = f(x(t), u(t)), \tag{4}$$

$x(t_0)$ given, where $x(t)$ is n dimensional and $u(t)$ is r dimensional. Under appropriate smoothness conditions it follows (see Section 9.5) that the gradient of J (with respect to u) is given by the function

$$l_u(x(t), u(t)) + \lambda'(t) f_u(x(t), u(t)), \tag{5}$$

where the $x(t)$ resulting from (4) when using the given control u is substituted in (5). The function $\lambda(t)$ is the solution of

$$-\dot{\lambda}(t) = f_x'(x, u)\lambda(t) + l_x'(x, u), \qquad \lambda(t_1) = \theta. \tag{6}$$

Thus, in summary, given a control function u the corresponding gradient of J can be found by integrating (4) forward to find x, then integrating (6) backward to find λ, and finally substituting the results in (5). This technique for calculating the gradient followed by the standard steepest-descent procedure is one of the most practical and efficient methods for solving unconstrained control problems.

CONJUGATE DIRECTION METHODS

10.6 Fourier Series

The problem of minimizing a quadratic functional on a Hilbert space can, by an appropriate transformation, be formulated as a Hilbert space minimum norm problem. It is then natural to look to the machinery of orthogonalization, the Gram-Schmidt procedure, and Fourier series to obtain a solution. This philosophy underlies conjugate direction methods.

Consider the quadratic objective functional f, defined on a Hilbert space H,

$$f(x) = (x \mid Qx) - 2(x \mid b),$$

where Q is a self-adjoint linear operator satisfying

$$\begin{aligned} (x \mid Qx) &\le M(x \mid x) \\ (x \mid Qx) &\ge m(x \mid x) \end{aligned} \tag{1}$$

for all $x \in H$ and some $M, m > 0$. Under these conditions the unique vector x_0 minimizing f is the unique solution of the equation $Qx = b$.

We can view this problem as a minimum norm problem by introducing the new inner product

$$[x \mid y] = (x \mid Qy),$$

since the problem is then equivalent to minimizing

$$\|x - x_0\|_Q^2 \equiv (x - x_0 \mid Q(x - x_0)).$$

Suppose we have, or can generate, a sequence of vectors $\{p_1, p_2, \ldots\}$ that are orthogonal with respect to the inner product $[\,|\,]$. Such a sequence is said to be *Q-orthogonal* or to be a sequence of *conjugate directions*. The vector x_0 can be expanded in a Fourier series with respect to this sequence. If the n-th partial sum of such an expansion is denoted x_n, then we have, by the fundamental approximation property of Fourier series, that $\|x_n - x_0\|_Q$ is minimized over the subspace $[p_1, p_2, \ldots, p_n]$. Therefore, as n increases, the value of $\|x_n - x_0\|_Q$ and the value of f decrease. If the sequence $\{p_i\}$ is complete, the process converges to x_0.

Of course, to compute the Fourier series of x_0 with respect to the $\{p_i\}$, we must be able to compute inner products of the form $[p_i \mid x_0]$. These are computable even though x_0 is unknown since $[p_i \mid x_0] = (p_i \mid Qx_0) = (p_i \mid b)$.

Theorem 1. (*Method of Conjugate Directions*) *Let $\{p_1\}$ be a sequence in H such that $(p_i \mid Qp_j) = 0$, $i \neq j$, and such that the closed linear subspace generated by the sequence is H. Then for any $x_1 \in H$ the sequence generated by the recursion.*

$$x_{n+1} = x_n + \alpha_n p_n \tag{2}$$

$$\alpha_n = \frac{(p_n \mid r_n)}{(p_n \mid Qp_n)} \tag{3}$$

$$r_n = b - Qx_n \tag{4}$$

satisfies $(r_n \mid p_k) = 0, k = 1, 2, \ldots, n-1$, and $x_n \to x_0$—the unique solution of $Qx = b$.

Proof. Define $y_n = x_n - x_1$. The recursion is then equivalent to $y_1 = \theta$ and

$$y_{n+1} = y_n + \frac{(p_n \mid b - Qx_1 - Qy_n)}{(p_n \mid Qp_n)} p_n, \tag{5}$$

or in terms of the inner product $[\,|\,]$,

$$y_{n+1} = y_n + \frac{[p_n \mid y_0 - y_n]}{[p_n \mid p_n]} p_n. \tag{6}$$

Since $y_n \in [p_1, p_2, \ldots, p_{n-1}]$ and since the p_i's are Q-orthogonal, it follows that $[p_n \mid y_n] = 0$ and hence equation (6) becomes

$$y_{n+1} = y_n + \frac{[p_n \mid y_0]}{[p_n \mid p_n]} p_n.$$

Thus

$$y_{n+1} = \sum_{k=1}^{n} \frac{[p_k \mid y_0]}{[p_k \mid p_k]} p_k,$$

which is the n-th partial sum of a Fourier expansion of y_0. Since with our assumptions on Q, convergence with respect to $\| \quad \|$ is equivalent to convergence with respect to $\| \quad \|_Q$, it follows that $y_n \to y_0$ and hence that $x_n \to x_0$.

The orthogonality relation $(r_n \mid p_k) = 0$ follows from the fact that the error $y_n - y_0 = x_n - x_0$ is Q-orthogonal to the subspace $[p_1, p_2, \ldots, p_{n-1}]$. ∎

Example 1. Consider once again the basic approximation problem in a Hilbert space X. We seek the vector

$$\hat{x} = \sum a_i y_i$$

in the subspace $[y_1, y_2, \ldots, y_n]$ which best approximates a given vector x. This leads to the normal equations

$$Ga = b,$$

where G is the Gram matrix of $\{y_1, y_2, \ldots, y_n\}$ and b is the vector with components $b_i = (x \mid y_i)$.

The n-dimensional linear equation is equivalent to the unconstrained minimization of

$$a'Ga - 2a'b$$

with respect to $a \in E^n$. This problem can be solved by using the method of conjugate directions. A set of linearly independent vectors $\{p_1, p_2, \ldots, p_n\}$ satisfying $p_i'Gp_j = 0$, $i \neq j$, can be constructed by applying the Gram-Schmidt procedure to any independent set of n vectors in E^n. For instance, we may take the vectors $e_i = (0, \ldots, 1, 0, \ldots, 0)$ (with the 1 in the i-th component) and orthogonalize these with respect to G. The resulting iterative calculation for a has as its k-th approximation a_k the vector such that

$$\hat{x}_k = \sum_{i=1}^{n} a_i^k y_i$$

is the best approximation to x in the subspace $[y_1, y_2, \ldots, y_k]$. In other words, this conjugate direction method is equivalent to solving the original

approximation problem by a Gram-Schmidt orthogonalization of the vectors $\{y_1, y_2, \ldots, y_n\}$.

*10.7 Orthogonalization of Moments

The Q-orthogonal direction vectors for a conjugate direction method can be obtained by applying the Gram-Schmidt procedure to any sequence of vectors that generate a dense subspace of H. Thus, if $\{e_i\}$ is such a sequence in H, we define

$$p_1 = e_1$$

$$p_n = e_n - \sum_{i=1}^{n-1} \frac{[e_i \mid p_i]}{[p_i \mid p_i]} p_i \qquad (n > 1),$$

where again $[x \mid y] \equiv (x \mid Qy)$. This procedure, in its most general form, is in practice rarely worth the effort involved. However, the scheme retains some of its attractiveness if, as practical considerations dictate, the sequence $\{e_n\}$ is not completely arbitrary but is itself generated by a simple recurrence scheme. Suppose, in particular, that starting with an initial vector e_1 and a bounded linear self-adjoint operator B, the sequence $\{e_i\}$ is generated by the relation $e_{n+1} = Be_n$. Such a sequence is said to be a sequence of *moments* of B. There appear to be no simple conditions guaranteeing that the moments generate a dense subspace of H, so we ignore this question here. The point of main interest is that a sequence of moments can be orthogonalized by a procedure that is far simpler than the general Gram-Schmidt procedure.

Theorem 1. *Let $\{e_i\}$ be a sequence of moments of a self-adjoint operator B. Then the sequence*

$$p_1 = e_1$$

$$p_2 = Bp_1 - \frac{[p_1 \mid Bp_1]}{[p_1 \mid p_1]} p_1$$

$$p_{n+1} = Bp_n - \frac{[p_n \mid Bp_n]}{[p_n \mid p_n]} p_n - \frac{[p_{n-1} \mid Bp_n]}{[p_{n-1} \mid p_{n-1}]} p_{n-1} \qquad (n \geq 2)$$

defines a Q-orthogonal sequence in H such that for each n, $[p_1, p_2, \ldots, p_n] = [e_1, e_2, \ldots, e_n]$.

Proof. Simple direct verification shows the theorem is true for p_1, p_2. We prove it for $n > 2$ by induction. Assume that the result is true for $\{p_i\}_{i=1}^{n}$. We prove that it is true for $\{p_i\}_{i=1}^{n+1}$.

It is clear by inspection that p_{n+1} is nonzero and is in the subspace $[e_1, e_2, \ldots, e_{n+1}]$. Therefore, it is only necessary to establish that p_{n+1} is orthogonal to each p_i, $i \leq n$. For any $i \leq n$ we have

$$[p_i \mid p_{n+1}] = [p_i \mid Bp_n] - \frac{[p_n \mid Bp_n]}{[p_n \mid p_n]}[p_i \mid p_n] - \frac{[p_n \mid Bp_{n-1}]}{[p_{n-1} \mid p_{n-1}]}[p_i \mid p_{n-1}].$$

For $i \leq n-2$ the second two terms in the above expression are zero by the induction hypothesis and the first term can be written as $[Bp_i \mid p_n]$ which is zero since Bp_i lies in the subspace $[p_1, p_2, \ldots, p_{i+1}]$. For $i = n-1$ the first and the third term cancel while the second term vanishes. For $i = n$ the first and the second term cancel while the third term vanishes. ∎

10.8 The Conjugate Gradient Method

A particularly attractive method of selecting direction vectors when minimizing the functional

$$f(x) = (x \mid Qx) - 2(b \mid x)$$

is to choose $p_1 = r_1 = b - Qx_1$ (the direction of the negative gradient of f at x_1) and then, after moving in this direction to x_2, consider the new negative gradient direction $r_2 = b - Qx_2$ and choose p_2 to be in the space spanned by r_1, r_2 but Q-orthogonal to p_1. We continue by selecting the other p_i's in a similar way. In other words, the sequence of p_i's is a Q-orthogonalized version of the sequence of negative gradients $\{r_1, r_2, \ldots\}$ generated as the descent process progresses. The method leads to the simple recursive form

$$x_{n+1} = x_n + \frac{(r_n \mid p_n)}{(p_n \mid Qp_n)} p_n \tag{1}$$

$$p_{n+1} = r_{n+1} - \frac{(r_{n+1} \mid Qp_n)}{(p_n \mid Qp_n)} p_n. \tag{2}$$

This two-term formula for the next member of the Q-orthogonal sequence $\{p_i\}$ can be considered a consequence of the theorem in the last section on orthogonalized moments. In the present case it is easily seen that r_{n+1} is in the subspace $[r_1, Qr_1, \ldots, Q^n r_1]$. Furthermore, because the direction vectors $\{p_n\}$ are generated from the negative gradients, the resulting closed subspace generated by them is always large enough so that the x_n's converge to the optimal solution.

Theorem 1. Let $x_1 \in H$ be given. Define $p_1 = b - Qx_1$ and

$$r_n = b - Qx_n \tag{3}$$

$$x_{n+1} = x_n + \alpha_n p_n \tag{4}$$

$$p_{n+1} = r_{n+1} - \beta_n p_n \tag{5}$$

$$\alpha_n = \frac{(r_n \mid p_n)}{(p_n \mid Qp_n)} \tag{6}$$

$$\beta_n = \frac{(r_{n+1} \mid Qp_n)}{(p_n \mid Qp_n)}. \tag{7}$$

Then the sequence $\{x_n\}$ converges to $x_0 = Q^{-1}b$.

Proof. We first show that this is a method of conjugate directions. Assume that this assertion is true for $\{p_k\}_{k=1}^{n}$, $\{x_k\}_{k=1}^{n+1}$; we shall show that it is true for one more step.

Since α_{n+1} is chosen in accordance with a method of conjugate directions we must only show that p_{n+1} is Q-orthogonal to the previous direction vectors. We have from (5)

$$(p_k \mid Qp_{n+1}) = (p_k \mid Qr_{n+1}) - \beta_n(p_k \mid Qp_n). \tag{8}$$

For $k = n$ the two terms on the right of (8) cancel. For $k < n$ the second term on the right is zero and the first term can be written as $(Qp_k \mid r_{n+1})$. But $Qp_k \in [p_1, p_2, \ldots, p_{k+1}] \subset [p_1, p_2, \ldots, p_n]$ and for any conjugate direction method $(r_{n+1} \mid p_i) = 0$, $i \leq n$. Hence the method is a conjugate direction method.

Next we prove that the sequence $\{x_n\}$ converges to x_0. Define the functional E by

$$E(x) = (b - Qx \mid Q^{-1}(b - Qx)).$$

We have, by direct evaluation,

$$E(x_n) - E(x_{n+1}) = \alpha_n(r_n \mid p_n).$$

But by (5) and $(r_n \mid p_{n-1}) = 0$ we have $(r_n \mid p_n) = (r_n \mid r_n)$ and hence

$$E(x_n) - E(x_{n+1}) = \alpha_n \frac{(r_n \mid r_n)}{(r_n \mid Q^{-1}r_n)} E(x_n). \tag{9}$$

Now from (5) and the Q-orthogonality of p_n and p_{n-1} we have

$$\begin{aligned}(r_n \mid Qr_n) &= (p_n \mid Qp_n) + \beta_{n-1}^2(p_{n-1} \mid Qp_{n-1}) \\ &\geq (p_n \mid Qp_n).\end{aligned} \tag{10}$$

Also, by definition of m, Section 10.6,

$$\frac{(r_n \mid r_n)}{(r_n \mid Q^{-1}r_n)} \geq m. \tag{11}$$

Hence, combining (9), (10), and (11), we obtain

$$E(x_{n+1}) \leq \left(1 - \frac{m}{M}\right) E(x_n).$$

Thus $E(x_n) \to 0$ which implies $r_n \to \theta$. ∎

The slight increase in the amount of computation required for the conjugate gradient method over that of steepest descent can lead to a significant improvement in the rate of convergence. It can be shown that for $m(x \mid x) \leq (x \mid Qx) \leq M(x \mid x)$ the convergence rate is

$$\|x_{n+1} - x_0\|^2 \leq \frac{4}{m} E(x_1) \left(\frac{1 - \sqrt{c}}{1 + \sqrt{c}}\right)^{2n},$$

where $c = m/M$, whereas for steepest descent the best estimate (see Problem 11) is

$$\|x_{n+1} - x_0\|^2 \leq \frac{1}{m} E(x_1) \left(\frac{1 - c}{1 + c}\right)^{2n}$$

In an n-dimensional quadratic problem the error tends to zero geometrically with steepest descent, in one step with Newton's method, and within n steps with any conjugate direction method.

The method of conjugate directions has several extensions applicable to the minimization of a nonquadratic functional f. One such method, the method of parallel tangents (PARTAN), is based on the easily established geometric relation which exists among the direction vectors for the quadratic version of the conjugate gradient method (see Figure 10.7).

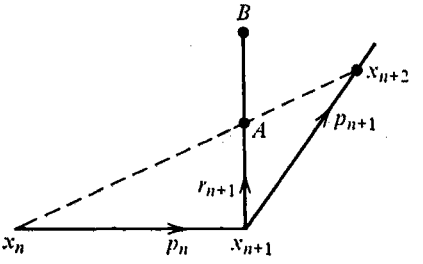

Figure 10.7 PARTAN

Point A, the intersection of the line between (x_n, x_{n+2}) with the line (x_{n+1}, B) determined by the negative gradient of f at x_{n+1}, is actually the minimum of f along the line (x_{n+1}, B). To carry out the PARTAN procedure for minimizing an arbitrary functional f, the point x_{n+2} is found from x_n and x_{n+1} by first minimizing f along the negative gradient direction from

x_{n+1} to find the point A and then minimizing f along the (dotted) line determined by x_n and A to find x_{n+2}. For a quadratic functional this method coincides with the conjugate gradient method. For nonquadratic functionals the process determines a decreasing sequence of functional values $f(x_n)$; practical experience indicates that it converges rapidly, although no sharp theoretical results are available.

METHODS FOR SOLVING CONSTRAINED PROBLEMS

Devising computational procedures for constrained optimization problems, as with solving any difficult problem, generally requires a lot of ingenuity and thorough familiarity with the basic principles and existing techniques of the area. No general, all-purpose optimization algorithm has been devised, but a number of procedures are effective for certain classes of problems. Essentially all of these methods have strong connections with the general principles discussed in the earlier chapters.

10.9 Projection Methods

One of the most common techniques for handling constraints is to use a descent method in which the direction of descent is chosen to decrease the cost functional and to remain within the constraint region.

The simplest version of the method is designed for problems of the form

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } Ax = b, \end{aligned}$$

where f is a functional on the Hilbert space X, A is a bounded linear operator from H into a Hilbert space Y, and b is fixed in Y. We assume that A has closed range. The procedure begins by starting from a point x_1 satisfying the constraint. An ideal direction vector p_1 is found by some standard technique such as steepest descent or Newton's method. This vector is projected onto the nullspace of A, $\mathcal{N}(A)$, giving the new direction vector g_1. The next point x_2 is then taken as

$$x_2 = x_1 + \alpha_1 g_1,$$

where α_1 is chosen in the usual way to minimize $f(x_2)$. Since $g_1 \in \mathcal{N}(A)$, the point x_2 also satisfies the constraint and the process can be continued.

To project the negative gradient at the n-th step onto $\mathcal{N}(A)$, the component of $f'(x_n)$ in $\mathcal{R}(A)$ must be added to the negative gradient. Thus the required projected negative gradient has the form

$$g_n = -f'(x_n) + A^*\lambda_n$$

and λ_n is chosen so that $A\{f'(x_n) - A^*\lambda_n\} = \theta$. Therefore,

$$g_n = -[I - A^*(AA^*)^{-1}A]f'(x_n).$$

At the solution x_0, $f'(x_0)$ will be orthogonal to $\mathcal{N}(A)$ and hence

$$f'(x_0) - A^*\lambda_0 = \theta.$$

This last equation is the necessary condition in terms of the Lagrange multiplier. Thus, as the method progresses, the computation of the projection of the gradient gives, in the limit, the Lagrange multiplier for the problem.

Similar considerations apply to nonlinear constraints of the form $H(x) = \theta$ and to inequality constraints of the form $G(x) \leq \theta$, but a number of additional calculations are required at each step. We do not discuss these techniques in detail here but merely indicate the general idea of one possible method. Additional information is contained in the references and the problems.

Suppose we seek to

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } H(x) = \theta, \end{aligned}$$

and that we have a point x_1 satisfying the constraint. To obtain an improved vector x_2, we project the negative gradient of f onto the tangent space $\{x : H'(x_1)x = \theta\}$ obtaining the direction vector g_1. Then $x_2^{(1)}$ is taken as $x_1 + \alpha_1 g_1$ where α_1 is chosen to minimize $f(x_2^{(1)})$. This new vector $x_2^{(1)}$ may not satisfy the constraint so it must be modified. One way is to employ a successive approximation technique to generate a sequence $\{x_2^{(k)}\}$ originating at $x_2^{(1)}$, which converges to a vector x_2 satisfying $H(x_2) = \theta$. Of course, once x_2 is finally obtained, it must be verified that $f(x_2) \leq f(x_1)$ so that x_2 is indeed an improvement. If $f(x_2) > f(x_1)$, α_1 must be reduced and a new x_2 found. The method is illustrated in Figure 10.8.

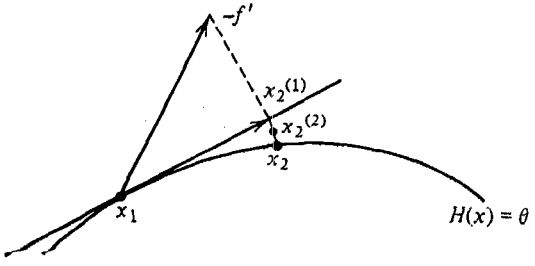

Figure 10.8 Gradient projection

There are a number of techniques of successive approximation that can be employed to move from the tangent plane to the constraint surface. The most common methods are modifications of Newton's method (see Problem 9).

10.10 The Primal-Dual Method

Duality theorems supply the basis for a number of computational procedures. These procedures, like the duality theorems themselves, often can be justified only for convex problems, but in many cases the basic idea can be modified so as to be effective for other problems.

We consider a dual method for the convex problem

$$\begin{cases}\text{minimize } f(x)\\ \text{subject to } G(x) \leq \theta, \qquad x \in \Omega.\end{cases} \tag{1}$$

Assuming that the constraint is regular, this problem is equivalent to

$$\max_{z^* \geq \theta} \inf_{x \in \Omega} \{f(x) + \langle G(x), z^* \rangle\}. \tag{2}$$

Or, defining the dual functional

$$\varphi(z^*) = \inf_{x \in \Omega} \{f(x) + \langle G(x), z^* \rangle\}, \tag{3}$$

the problem is equivalent to the dual problem

$$\begin{cases}\text{maximize } \varphi(z^*)\\ \text{subject to } z^* \geq \theta.\end{cases} \tag{4}$$

The dual problem (4) has only the constraint $z^* \geq \theta$, hence, assuming that the gradient of φ is available, the dual problem can be solved in a rather routine fashion. (Note that if the primal problem (1) has only equality constraints of the form $Ax = b$, the dual problem (4) will have no constraints.) Once the dual problem is solved yielding an optimal z_0^*, the primal problem can be solved by minimizing the corresponding Lagrangian.

Example 1. (Hildreth's Quadratic Programming Procedure) Consider the constrained quadratic minimization problem:

$$\begin{cases}\text{minimize } \frac{1}{2}x'Qx - b'x\\ \text{subject to } Ax \leq c.\end{cases} \tag{5}$$

Here x is an n vector, Q an $n \times n$ positive-definite matrix, A an $m \times n$ matrix, and b and c are given vectors of dimension n and m, respectively.

The dual problem (see Example 1, Section 8.6) is

$$\begin{aligned} &\text{maximize } -\tfrac{1}{2}\lambda' P\lambda - \lambda' d - \tfrac{1}{2}b'Q^{-1}b \\ &\text{subject to } \lambda \geq \theta, \end{aligned}$$

where $P = AQ^{-1}A'$, $d = AQ^{-1}b + c$. Or, equivalently,

$$\begin{cases} \text{minimize } \tfrac{1}{2}\lambda' P\lambda + \lambda' d \\ \text{subject to } \lambda \geq \theta. \end{cases} \tag{6}$$

After solving Problem (6), obtaining λ_0, the solution to (5) is

$$x_0 = Q^{-1}(b - A'\lambda_0).$$

To solve (6) we employ a descent procedure with direction vectors equal to the usual basis vectors $e_i = (0, 0, \ldots, 1, 0, \ldots, 0)$. Specifically we let the infinite sequence of direction vectors be $\{e_1, e_2, \ldots, e_n, e_1, e_2, \ldots, e_n, \ldots\}$. Thus we vary the vector λ one component at a time.

At a given step in the process, having obtained a vector $\lambda \geq \theta$, we fix our attention on a single component λ_i. The objective functional may be regarded as a quadratic function of this one component. We adjust λ_i to minimize the function, or if that would require $\lambda_i < 0$, we set $\lambda_i = 0$. In any case, however, the objective functional is decreased. Then we consider the next component λ_{i+1}.

If we consider one complete cycle through the components to be one iteration taking the vector λ^k to λ^{k+1}, the method can be expressed explicitly as

$$\lambda_i^{k+1} = \max(0, w_i^{k+1}),$$

where

$$w_i^{k+1} = -\frac{1}{p_{ii}}\left(d_i + \sum_{j=1}^{i-1} p_{ij}\lambda_j^{k+1} + \sum_{j=i+1}^{n} p_{ij}\lambda_j^k\right).$$

Convergence of the method is easily proved.

Although the dual functional can be evaluated analytically in only a few special cases, solving the dual rather than the primal is often an efficient procedure. Each evaluation of the dual functional, however, requires solving an unconstrained minimization problem.

An evaluation of the dual functional by minimization yields the gradient of the dual as well as its value. Suppose that x_1 is a minimizing vector in (3) corresponding to z_1^*. Then for arbitrary z^*

$$\varphi(z_1^*) = f(x_1) + \langle G(x_1), z_1^* \rangle$$

$$\varphi(z^*) \leq f(x_1) + \langle G(x_1), z^* \rangle.$$

Therefore,

$$\varphi(z^*) - \varphi(z_1^*) \leq \langle G(x_1), z^* - z_1^* \rangle \tag{7}$$

and hence $G(x_1)$ defines a hyperplane that bounds φ from above. If φ is differentiable, $G(x_1)$ is the (unique) gradient of φ at z_1^*. (See Problem 8, Chapter 8.)

In view of the above observation, it is not difficult to solve the dual problem by using some gradient-based technique modified slightly to account for the constraint $z^* \geq \theta$. At each step of the process an unconstrained minimization is performed with respect to x in order to evaluate φ and its gradient. The minimization with respect to x must, of course, usually be performed by some iterative technique.

Example 2. (Optimal Control) Suppose a dynamic system is governed by an n-th order set of differential equations

$$\dot{x}(t) = f(x, u, t).$$

Given an initial state $x(t_0)$, we consider the problem of selecting the m-dimensional control $u(t)$, $t_0 \leq t \leq t_1$, such that $u(t) \in U \subset R^m$ so as to minimize the convex functional

$$\psi(x(t_1))$$

subject to the terminal constraints

$$G(x(t_1)) \leq \theta,$$

where G is a convex mapping of E^n into E^r.

Let us partition the state vector into two parts, $x = (y, z)$, where y is that part of the state vector that enters explicitly into the cost functional ψ and the constraints G. We write $\psi(y)$, $G(y)$ for $\psi(x)$, $G(x)$, respectively. For many problems the dimension p of y is equal to $m + 1$. For example, in optimizing the flight of a rocket to a given target, the components in $x(t_1)$ representing velocity do not explicitly enter the terminal position constraint.

To apply the primal-dual algorithm to this problem, we define the set

$\Gamma = \{y \in E^p : (y, z) = x(t_1)$, where $x(t_1)$ is the terminal point of some trajectory generated by a feasible control input $u\}$.

The control problem is then equivalent to the finite-dimensional problem

$$\begin{aligned} &\text{minimize } \psi(y) \\ &\text{subject to } G(y) \leq \theta, \qquad y \in \Gamma. \end{aligned}$$

This is a convex programming problem if Γ is convex. For many nonlinear systems, Γ can be argued to be convex if the dimension p of y is sufficiently low. By duality the problem is then equivalent to the problem

$$\max_{\lambda \geq 0} \left\{ \min_{y \in \Gamma} [\psi(y) + \lambda' G(y)] \right\}.$$

For fixed λ the inner minimization over y is equivalent to an optimal control problem having terminal cost

$$\psi(x(t_1)) + \lambda' G(x(t_1))$$

but having no terminal constraints. This latter type of control problem can be solved by a standard gradient method (see Example 2, Section 10.5).

10.11 Penalty Functions

It has long been common practice among optimizers to attack a problem such as

$$(1) \qquad \begin{cases} \text{minimize } f(x) \\ \text{subject to } h_i(x) = 0, \qquad i = 1, 2, \ldots, p, \end{cases}$$

by solving instead the unconstrained approximating problem

$$(2) \qquad \text{minimize } f(x) + K \sum_i h_i^2(x)$$

for some large positive constant K. For sufficiently large K it can be reasoned that the solutions to problems (1) and (2) will be nearly equal. The term $K \sum_i h_i^2(x)$ is referred to as a penalty function since in effect it assigns a specific cost to violations of the constraints.

In the practical implementation of the penalty function method, we are driven on the one hand to select K as large as possible to enhance the degree of approximation, and on the other to keep K somewhat small so that when calculating gradients the penalty terms do not completely swamp out the original objective functional. A common technique is to progressively solve problem (2), the unconstrained approximation, for a sequence of K's which tend toward infinity. The resulting sequence of approximate solutions can then be expected to converge to the solution of the original constrained problem. In this section we investigate this type of scheme as applied to inequality as well as equality constraints.

At first the penalty function method may appear to be a simple algebraic device—a somewhat crude scheme for overcoming the difficulties imposed by constraints. There is a geometric interpretation of the method, however,

which illuminates its intimate relation with other optimization techniques and lends a degree of elegance to the scheme. Problem (1) is clearly equivalent to

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \sum_i h_i^2(x) \leq 0, \end{aligned} \tag{3}$$

and by this transformation we reduce the n constraints to a single constraint. This constraint, it should be noted, is not regular; i.e., there is no x such that $\sum_i h_i^2(x) < 0$. The primal function for problem (3),

$$\omega(z) = \inf \{f(x) : \sum h_i^2(x) \leq z\},$$

looks something like that shown in Figure 10.9. It is nonincreasing with $z = 0$ as a boundary point of its region of definition. The hyperplane (which in this case is merely a line since z is a real variable) supporting the shaded region at the point $(\omega(0), 0)$ may be vertical.

Specifying $K > 0$ and minimizing $f(x) + K\sum_i h_i^2(x)$ determines, as shown in Figure 10.9, a supporting hyperplane and a value φ_K for the dual functional corresponding to problem (3). Provided ω is continuous, it is clear that as K is increased, φ_K will increase monotonically toward $\omega(0)$. No convexity requirements need be imposed; since 0 is a boundary point of the region of definition for ω, a (perhaps vertical) support hyperplane always exists.

There are numerous variations of this scheme. For instance, since the n original equality constraints are also equivalent to the single constraint

$$\sum_i |h_i(x)| \leq 0,$$

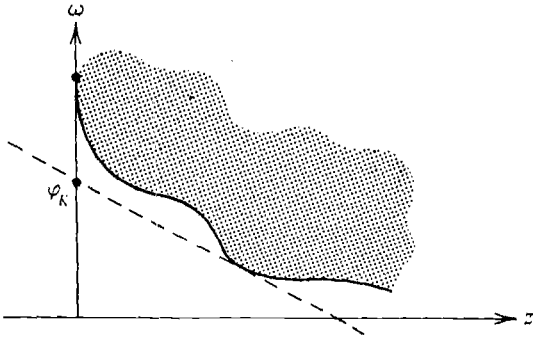

Figure 10.9 The primal function

the penalty term $K\sum|h_i(x)|$ can be used. If the h_i's have nonvanishing first derivatives at the solution, the primal function will, in this case, as in Figure 10.10, have finite slope at $z = 0$, and hence some finite value of K will yield a support hyperplane. This latter feature is attractive from a computational point of view but is usually offset by the difficulties imposed by nonexistence of a gradient of $\sum|h_i(x)|$.

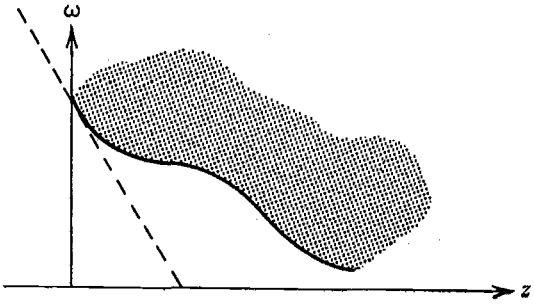

Figure 10.10 The primal function using $\sum|h_i|$

For inequality constraints a similar technique applies. Given a functional g on a vector space X, we define

$$g^+(x) = \max\{0, g(x)\}.$$

Let G be a mapping from X into R^p, i.e., $G(x) = (g_1(x), g_2(x), \ldots, g_p(x))$. We define

$$G^+(x) = (g_1{}^+(x), g_2{}^+(x), \ldots, g_p{}^+(x)).$$

It is then clear the p inequalities $G(x) \leq \theta$ are equivalent to the single inequality

$$G^+(x)'G^+(x) \equiv \sum_i [g_i{}^+(x)]^2 \leq 0.$$

Again this inequality does not satisfy the regularity condition. Since this form includes equality constraints, we consider only inequalities in the remainder of the section. Hence we analyze the method in detail for the problem

$$\text{(4)} \qquad \begin{cases}\text{minimize } f(x)\\ \text{subject to } G(x) \leq \theta.\end{cases}$$

The geometric interpretation of this problem is identical with that for equalities.

Throughout the following it is assumed that $\{K_n\}$ is an increasing

sequence of positive constants tending toward infinity and that problem (4) has a solution.

Lemma 1. *Define*

$$\mu_0 = \min \{f(x) : G(x) \leq \theta\}.$$

For each n let x_n be a point minimizing

$$f_n(x) \equiv f(x) + K_n G^+(x)'G^+(x).$$

Then

1. $f_{n+1}(x_{n+1}) \geq f_n(x_n)$.
2. $\mu_0 \geq f_n(x_n)$.
3. $\lim_{n \to \infty} K_n G^+(x_n)'G^+(x_n) = 0$.

Proof.

1. $$\begin{aligned} f_{n+1}(x_{n+1}) &= f(x_{n+1}) + K_{n+1} G^+(x_{n+1})'G^+(x_{n+1}) \\ &\geq f(x_{n+1}) + K_n G^+(x_{n+1})'G^+(x_{n+1}) \\ &\geq f_n(x_n). \end{aligned}$$

2. Let x_0 solve (4). Then $\mu_0 = f(x_0) = f_n(x_0) \geq f_n(x_n)$.
3. For convenience define $g(x) = G^+(x)'G^+(x)$. Since $K_n\, g(x) \geq 0$, it is only necessary to prove that $\limsup K_n\, g(x_n) = 0$. Assume to the contrary that $\limsup K_n\, g(x_n) = 3\varepsilon > 0$. Since by parts 1 and 2 the sequence $\{f_n(x_n)\}$ is nondecreasing and bounded, it therefore has a limit γ. Select N such that

$$\gamma < f_N(x_N) + \varepsilon.$$

Since $K_n \to \infty$ and since $\limsup K_n\, g(x_n) = 3\varepsilon$, there is an integer M such that

$$K_M > 4K_N$$

and

$$2\varepsilon < K_M\, g(x_M) < 4\varepsilon.$$

We then obtain a contradiction from the following string of inequalities:

$$\begin{aligned} \gamma &< f_N(x_N) + \varepsilon \leq f(x_M) + K_N g(x_M) + \varepsilon \\ &= f(x_M) + \left(\frac{K_N}{K_M}\right) K_M\, g(x_M) + \varepsilon < f(x_M) + \tfrac{1}{4} K_M\, g(x_M) + \varepsilon \\ &< f(x_M) + 2\varepsilon < f(x_M) + K_M\, g(x_M) = f_M(x_M) \leq \gamma. \end{aligned}$$

∎

Notice that part 3 of Lemma 1 is stronger than the statement

$$G^+(x_n)'G^+(x_n) \to 0.$$

Notice also that no continuity, convergence, or convexity assumptions are required by Lemma 1.

Theorem 1. *Let $f(x)$ and $G^+(x)'G^+(x)$ be lower semicontinuous functionals. If x_0 is any limit point of the sequence $\{x_n\}$ defined in Lemma 1, then x_0 solves problem (4).*

Proof. Since $g(x_n) \equiv G^+(x_n)'G^+(x_n) \to 0$ by Lemma 1, it follows by the lower semicontinuity of g that $g(x_0) = 0$ and hence $G(x_0) \leq \theta$. We have $\mu_0 \leq f(x_0)$ since $G(x_0) \leq \theta$. Also $f(x_n) \leq f_n(x_n) \leq \mu_0$ and hence by the lower semicontinuity of f, $f(x_0) \leq \mu_0$. ∎

It is remarkable and yet perhaps inevitable that the penalty function method, a method so simply conceived, is strongly connected with the theory of Lagrange multipliers. The connection is greatest for convex problems where, in fact, the penalty function method emerges as a particularly nice implementation of the primal-dual philosophy.

Lemma 2. *Let f and G be convex and continuous on the normed space X. If x_0 minimizes*

$$f(x) + G^+(x)'G^+(x), \tag{5}$$

then it also minimizes

$$f(x) + \lambda_0'G(x), \tag{6}$$

where $\lambda_0 = 2G^+(x_0)$.

Proof. The reader should find it simple but instructive to supply a proof assuming differentiability of f and G.

Without assuming differentiability, we construct a proof by contraposition. Suppose x_0 does not minimize (6); specifically suppose there is x_1 and $\varepsilon > 0$ such that

$$f(x_1) + \lambda_0'G(x_1) < f(x_0) + \lambda_0'G(x_0) - \varepsilon.$$

Let $x_\alpha = \alpha x_1 + (1-\alpha)x_0$. We write the identity (for $0 \leq \alpha \leq 1$)

$$\begin{aligned} f(x_\alpha) + G^+(x_\alpha)'G^+(x_\alpha) = {} & f(x_0) + G^+(x_0)'G^+(x_0) \\ & + f(x_\alpha) + \lambda_0'G(x_\alpha) - f(x_0) - \lambda_0'G(x_0) \\ & + \|G^+(x_\alpha) - G^+(x_0)\|^2 - \lambda_0'[G(x_\alpha) - G^+(x_\alpha)]. \end{aligned}$$

Because of the definition of x_1 the second line of the expression is less than $-\alpha\varepsilon$. Since a continuous convex functional satisfies a Lipschitz condition at

every point on the interior of its domain (see problem 19), the first term in the third line is $O(\alpha^2)$. The last term is identically zero for α sufficiently small. Thus

$$f(x_\alpha) + G^+(x_\alpha)'G^+(x_\alpha) \leq f(x_0) + G^+(x_0)'G^+(x_0) - \varepsilon\alpha + O(\alpha^2),$$

and hence the left side is less than the right for sufficiently small α showing that x_0 does not minimize (5). ∎

Suppose we apply the penalty function method to a problem where f and G are convex and continous. At the n-th step let x_n minimize

$$f(x) + K_n G^+(x)'G^+(x)$$

and define

$$\lambda_n = 2K_n G^+(x_n).$$

Then applying Lemma 2 with $G \to K_n^{1/2}G$, it follows that x_n also minimizes

$$f(x) + \lambda_n'G(x_n)$$

or, in terms of the dual functional φ corresponding to problem 4,

$$\varphi(\lambda_n) = f(x_n) + \lambda_n'G(x_n).$$

In other words, x_n, the result of the n-th penalty function minimization, determines a dual vector λ_n and the corresponding value of the dual functional. This leads to the interpretation that the penalty function method seeks to solve the dual.

Theorem 2. *Let f and G be convex and continuous. Suppose x_n minimizes $f(x) + K_n G^+(x)'G^+(x)$ and define $\lambda_n = 2K_n G^+(x_n)$. If λ_0 is any limit point of the sequence $\{\lambda_n\}$, then λ_0 solves the dual problem*

$$\begin{aligned} &\textit{maximize } \varphi(\lambda) \\ &\textit{subject to } \lambda \geq \theta. \end{aligned}$$

Proof. The dual functional φ is concave and, being a conjugate functional (except for sign), it is upper semicontinuous (see Sections 7.9 and 7.10). Hence $\varphi(\lambda_0) \geq \limsup \varphi(\lambda_n)$. Now given any $\lambda \geq \theta$ and any $\varepsilon > 0$, select N large enough so that for all $n \geq N$ we have $\lambda'G(x_n) < \varepsilon$. This choice is possible since $G^+(x_n) \to \theta$. Then, since $\lambda_n'G(x_n) \geq 0$, we have

$$\varphi(\lambda) \leq f(x_n) + \lambda'G(x_n) \leq f(x_n) + \lambda_n'G(x_n) + \varepsilon = \varphi(\lambda_n) + \varepsilon$$

for $n \geq N$. Therefore, since ε was arbitrary, we have $\varphi(\lambda) \leq \limsup \varphi(\lambda_n) \leq \varphi(\lambda_0)$. ∎

10.12 Problems

1. Let S be a closed subset of a Banach space. A mapping T from S onto a region Γ containing S is an *expansion mapping* if there is a constant $K > 1$ such that $\|T(x) - T(y)\| \geq K\|x - y\|$ for $x \neq y$. Show that an expansion mapping has a unique fixed point.
2. Let S be a compact subset of a Banach space X and let T be a mapping of S into S satisfying $\|T(x) - T(y)\| < \|x - y\|$ for $x \neq y$. Show that T has a unique fixed point in S which can be found by the method of successive approximation.
3. Let $X = L_2[a, b]$. Suppose the real-valued function f is such that for each $x \in X$

$$\int_a^b f(t, s, x(s))\, ds$$

is an element of X. Suppose also that $|f(t, s, \xi) - f(t, s, \xi')| \leq K(t, s)|\xi - \xi'|$, where $\int_a^b \int_a^b K(t, s)^2\, ds\, dt < 1$. Show that the integral equation

$$x(t) = y(t) + \int_a^b f(t, s, x(s))\, ds$$

has a unique solution $x \in X$ for every $y \in X$.
4. Let X be a Banach space and let A be a bounded linear operator from X into X. Using the contraction mapping theorem, show that if $\|A\| = a < 1$, then $(I - A)^{-1}$ exists (where I is the identity operator) and $\|(I - A)^{-1}\| < 1/(1 - a)$.
5. Using a technique similar to that employed for solving Volterra integral equations, devise a successive approximation scheme with guaranteed convergence for solving the two-point boundary value problem associated with minimizing

$$\int_{t_0}^{t_1} [x'(t)x(t) + u^2(t)]\, dt$$

subject to $\dot{x}(t) = f[x(t)] + bu(t)$, $x(t_0)$ fixed.
6. Show that Newton's method applied to a function f of the single real variable x converges monotonically after the second step if $f'(x) > 0$ and $f''(x) > 0$ everywhere.
7. In a modified Newton's method for solving $P(x) = \theta$, we iterate according to

$$x_{n+1} = x_n - [P'(x_1)]^{-1} P(x_n). \tag{1}$$

Assume that P is Fréchet differentiable in a convex region D and that for some $x_1 \in D$, $[P'(x_1)]^{-1}$ exists. Assume that x_2 calculated according to equation (1) is in D, that

$$\rho = \|[P'(x_1)]^{-1}\| \sup_{x \in D} \|P'(x_1) - P'(x)\| < 1,$$

and that the sphere

$$S = \{x : \|x - x_2\| < \frac{\rho}{1-\rho} \|x_1 - x_2\|\}$$

is contained in D. Show that the modified method converges to a solution $x_0 \in S$.

8. Use the modified Newton's method to calculate $\sqrt{10}$ starting from $x_1 = 3$ and from $x_1 = 1$.
9. Let X and Y be Hilbert spaces and let P be a transformation from X into Y. Suppose that P has a Fréchet derivative $P'(x)$ at each point $x \in X$ and that $P'(x)$ has closed range. Show that, under conditions similar to those of the standard Newton's method theorem, the sequence $x_{n+1} = x_n - [P'(x_n)]^{\dagger} P(x_n)$ converges to a point x_0 satisfying $P(x_0) = \theta$. ($A^{\dagger}$ denotes the pseudoinverse of A.)
10. Suppose the bounded, self-adjoint operator Q on a Hilbert space X satisfies

$$\inf \frac{(x \mid Qx)}{(x \mid x)} = m > 0.$$

Show that Q has an inverse and that for all x

$$(x \mid Q^{-1}x) \leq \frac{1}{m} (x \mid x).$$

11. Use Kantorovich's inequality

$$(x \mid Qx)(x \mid Q^{-1}x)/(x \mid x)^2 \leq (m+M)^2/4mM$$

to obtain an improved estimate for the convergence rate of steepest descent over that given in Theorem 1, Section 10.5. Prove the inequality for positive-definite (symmetric) matrices.

12. Let f be a functional defined on a normed space X and bounded below. Given $x_1 \in X$, let S be the closed convex hull of the set $\{x : f(x) < f(x_1)\}$. Assume that S is bounded and that f has a uniformly continuous Fréchet derivative on S. Show that the method of steepest descent applied to f from x_1 generates a sequence $\{x_n\}$ such that $\|f'(x_n)\| \to \theta$.

13. Under the hypothesis of Theorem 2, Section 10.5, show that the simplified steepest-descent process defined by $x_{n+1} = x_n - (1/M)f'(x_n)$ converges to the point minimizing f.
14. Suppose a sequence $\{x_n\}$ in a normed space converges to a point x_0. The convergence is said to be *weakly linear* if there exists a positive integer N such that

$$\limsup_{n\to\infty} \frac{\|x_{n+N} - x_0\|}{\|x_n - x_0\|} < 1.$$

(a) Show that in Theorem 1, Section 10.5, the convergence is weakly linear.
(b) Show that in Theorem 2, Section 10.5, the convergence is weakly linear.

15. Let B be a bounded linear operator mapping a Hilbert space H into itself and let e_1 be an arbitrary element of H. Let $\{e_k\}$ be the sequence of moments $e_{k+1} = Be_k$. Show that if $e_n \in [e_1, e_2, \ldots, e_{n-1}]$, then $e_m \in [e_1, e_2, \ldots, e_{n-1}]$ for all $m > n$.
16. Show that in the method of conjugate gradients there is the relation $(r_n \mid p_n) = (r_1 \mid p_n)$.
17. Verify that the geometric relation of Figure 10.7 holds for the method of conjugate gradients.
18. Show that even if f and G are not convex, if the primal-dual method converges in the sense that $z_n^* \to z_0^*$ where $z_0^* \geq \theta$, $\varphi'(z_0^*) \leq \theta$, and $\langle z_0^*, \varphi'(z_0^*)\rangle = 0$, then the corresponding x_0 minimizing $f(x) + \langle G(x), z_0^*\rangle$ is optimal.
19. Show that a continuous convex functional satisfies a Lipschitz condition at every point in the relative interior of its domain of definition. Hint: See the end of the Proof of Proposition 1, Section 7.9.
20. An alternative penalty function method for minimizing $f(x)$ over the constraint region

$$\Omega = \{x : g_i(x) \geq 0,\ i = 1, 2, \ldots, m\}$$

is to find a local minimum of

$$f(x) + r\sum_{i=1}^{m} \frac{1}{g_i(x)}$$

over the interior of the region Ω. For small r the solutions to the two problems will be nearly equal. Develop a geometric interpretation of this method.

REFERENCES

§10.2. The contraction mapping theorem is the simplest of a variety of fixed-point theorems. For an introduction to other results, see Courant and Robbins [32], Graves [64], and Bonsall [24].

§10.3. Much of the theory of Newton's method in Banach spaces was developed by Kantorovich [78]. See also Kantorovich and Akilov [79], [80], Bartle [16], Antosiewicz and Rheinboldt [9], and Collatz [30]. Some interesting extensions and modifications of the method have been developed by Altman [6], [7], [8]. An interesting account of various applications of Newton's method is found in Bellman and Kalaba [18].

§10.4–5. For material on steepest descent, see Ostrowski [115], Rosenbloom [130], Nashed [108], Curry [33], and Kelley [84].

§10.6–8. For the original development of conjugate direction methods, see Hestenes and Stiefel [71] and Hestenes [70]. The underlying principle of orthogonalizing moments, useful for a number of computational problems, is discussed in detail by Vorobyev [150] and Faddeev and Faddeeva [50]. There are various extensions of the method of conjugate gradients to problems with nonquadratic objective functionals. The most popular of these are the Fletcher and Reeves [54] method and the PARTAN method in Shah, Buehler, and Kempthorne [138]. For another generalization and the derivation of the convergence rate see Daniel [34], [35]. For application to optimal control theory, see Lasdon, Mitter, and Waren [94] and Sinnott and Luenberger [140].

§10.9–11. The gradient-projection method is due to Rosen [128], [129]. A closely related method is that of Zoutendijk [158]. For an application of the primal-dual method, see Wilson [154]. The penalty function method goes back to a suggestion of Courant [31]. See also Butler and Martin [27], Kelley [84], Fabian [49], and Fiacco and McCormick [53]. For additional general material, bibliography, and review of optimization techniques, see Saaty and Bram [136], Wolfe [156], Zoutendijk [159], Leon [96], Wilde and Beightler [153], Dorn [43], and Spang [143].

§10.12. For extensions of Problem 9 to Banach space, see Altman [6] and the generalized inverse function theorem of Chapter 9. For Kantorovich's inequal -ity (Problem 11), see Greub and Rheinboldt [66]. The method introduced in Problem 20 has been studied extensively by Fiacco and McCormick [53].

BIBLIOGRAPHY

[1] Abadie, J., ed., *Nonlinear Programming*, North-Holland Pub. Co., Amsterdam, 1967.

[2] Akhiezer, N. I., *The Calculus of Variations*, Ginn (Blaisdell), Boston, 1962.

[3] Akhiezer, N. I. and I. M. Glazman, *Theory of Linear Operators in Hilbert Space*, Vols. I, II, Frederick Ungar, New York, 1961.

[4] Akhiezer, N. I. and M. Krein, "Some Questions in the Theory of Moments," Article 4, *Amer. Math. Soc. Publ.*, 1962.

[5] Albert, A., "An Introduction and Beginner's Guide to Matrix Pseudo-Inverses," *ARCON*, July 1964.

[6] Altman, M., "A Generalization of Newton's Method," *Bull. Acad. Polon. Sci. Cl. III*, **3**, 189, 1955.

[7] Altman, M., "On the Generalization of Newton's Method," *Bull. Acad. Polon. Sci. Cl. III*, **5**, 789, 1957.

[8] Altman, M., "Connection between the Method of Steepest Descent and Newton's Method," *Bull. Acad. Polon. Sci. Cl. III*, **5**, 1031–1036, 1957.

[9] Antosiewicz, H. A. and W. C. Rheinboldt, "Numerical Analysis and Functional Analysis," Chapter 14 in *Survey of Numerical Analysis*, ed. by J. Todd, McGraw-Hill, New York, 1962.

[10] Apostol, T. M., *Mathematical Analysis, A Modern Approach to Advanced Calculus*, Addison-Wesley, Reading, Mass., 1957.

[11] Aronszajn, N., "Theory of Reproducing Kernels," *Trans. Amer. Math. Soc.*, **68**, 337–404, 1950.

[12] Arrow, K. J., L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-Linear Programming*, Chapter 4, "Programming in Linear Spaces," Stanford Univ. Press, Stanford, Calif., 1964, pp. 38–102.

[13] Arrow, K. J., S. Karlin, and H. Scarf, *Studies in the Mathematical Theory of Inventory and Production*, Stanford Univ. Press, Stanford, Calif., 1958.

[14] Balakrishnan, A. V., "An Operator-Theoretic Formulation of a Class of Control Problems and a Steepest Descent Method of Solution, *J. SIAM Control*, Ser. A, **1** (2), 1963.

[15] Balakrishnan, A. V., "Optimal Control Problems in Banach Spaces," *J. SIAM Control*, Ser. A, **3** (1), 152–180, 1965.

[16] Bartle, R. G., "Newton's Method in Banach Spaces," *Proc. Amer. Math. Soc.*, **6**, 827–831, 1955.

[17] Bellman, R. E., I. Glicksberg, and O. A. Gross, *Some Aspects of the Mathematical Theory of Control Processes*, The Rand Corp., Santa Monica, Calif., Jan. 16, 1958.

[18] Bellman, R. E. and R. E. Kalaba, *Quasilinearization and Nonlinear Boundary-Value Problems*, American Elsevier, New York, 1965.

[19] Bellman, R. E. and W. Karush, "Mathematical Programming and the Maximum Transform," *J. Soc. Indust. Appl. Math.*, **10**, 550–567, 1962.

[20] Ben-Israel, A. and A. Charnes, "Contributions to the Theory of Generalized Inverses," *J. Soc. Indust. Appl. Math.*, **11** (3), Sept. 1963.

[21] Berberian, S. K., *Introduction to Hilbert Space*, Oxford Univ. Press, 1961.

[22] Bliss, G. A., *Lectures on the Calculus of Variations*, Univ. of Chicago Press, 1945.

[23] Blum, E. K., "The Calculus of Variations, Functional Analysis, and Optimal Control Problems," *Topics in Optimization*, Ed. by G. Leitman, Academic Press, New York, 417–461, 1967.

[24] Bonsall, F. F., *Lectures on Some Fixed Point Theorems of Functional Analysis*, Tata Inst. of Fundamental Res., Bombay, 1962.

[25] Brondsted, A. "Conjugate Convex Functions in Topological Vector Spaces," *Mat. Fys. Medd. Dan. Vid. Selsk.* **34** (2), 1–27, 1964.

[26] Butkovskii, A. G., "The Method of Moments in the Theory of Optimal Control of Systems with Distributed Parameters," *Automation and Remote Control*, **24**, 1106–1113, 1963.

[27] Butler, T. and A. V. Martin, "On a Method of Courant for Minimizing Functionals," *J. Math. and Physics*, **41**, 291–299, 1962.

[28] Canon, M., C. Cullum, and E. Polak, "Constrained Minimization Problems in Finite-Dimensional Spaces," *J. SIAM Control*, **4** (3), 528–547, 1966.

[29] Chipman, J. S., "On Least Squares with Insufficient Observations," *J. Amer. Stat. Assoc.*, **59**, 1078–1111, Dec. 1964.

[30] Collatz, L., *Functional Analysis and Numerical Mathematics*, trans. by H. Oser, Academic Press, New York, 1966.

[31] Courant, R., "Calculus of Variations and Supplementary Notes and Exercises" (mimeographed notes), Supplementary notes by M. Kruskal and H. Rubin, revised and amended by J. Moser, New York Univ., 1962.

[32] Courant, R. and H. Robbins, *What is Mathematics?* Oxford Univ. Press, London, 1941, pp. 251–255.

[33] Curry, H. B., "The Method of Steepest Descent for Nonlinear Minimization Problems," *Quar. Appl. Math.*, **2**, 258–261, 1944.

[34] Daniel, J. W., "The Conjugate Gradient Method for Linear and Nonlinear Operator Equations," *J. SIAM Numer. Anal.*, **4** (1), 10–25, 1967.

[35] Daniel, J. W., "Convergence of the Conjugate Gradient Method with Computationally Convenient Modifications," *Numerische Mathematik*, **10**, 125–131, 1967.

[36] Davis, P. J., *Interpolation and Approximation*, Blaisdell, New York, 1965.

[37] Day, M. M., *Normed Linear Spaces*, Springer, Berlin, 1958.

[38] Desoer, C. A. and B. H. Whalen, "A Note on Pseudoinverses," *J. Soc. Indust. Appl. Math.*, **11** (2), 442–447, June 1963.

[39] Deutsch, F. R. and P. H. Maserick, "Applications of the Hahn-Banach Theorem in Approximation Theory," *SIAM Rev.*, **9** (3), 516–530, July 1967.

[81] Karlin, S. "Operator Treatment of Minimax Principle," *Contributions to the Theory of Games*, ed. by H. W. Kuhn and A. W. Tucker, Princeton Univ. Press, 1950, pp. 133–154.

[82] Karlin, S. *Mathematical Methods and Theory in Games, Programming, and Economics*, Vol. I, Addison-Wesley, Reading Mass., 1959.

[83] Karlin, *S. Mathematical Methods and Theory in Games, Programming, and Economics*, Vol. II, Addison-Wesley, Reading, Mass., 1959.

[84] Kelley, H. J., "Method of Gradients," Chapter 6 of *Optimization Techniques*, ed. by G. Leitmann, Academic Press, New York, 1962, pp. 206–252.

[85] Kelley, J. L. and I. Namioka, et al., *Linear Topological Spaces*, Van Nostrand, Princeton, N.J., 1963.

[86] Kirillova, F. M., "Applications of Functional Analysis to the Theory of Optimal Processes," *J. SIAM Control*, **5** (1), 25–50, Feb. 1967.

[87] Klee, V. L., "Separation and Support Properties of Convex Sets," *Seminar on Convex Sets*, The Institute for Advanced Study, Princeton, N.J., 1949–1950, pp. 77–87.

[88] Kolmogorov, A. N. and S.V. Fomin, *Elements of the Theory of Functional Analysis*, Vol. 1, Graylock Press, Rochester, N.Y., 1957.

[89] Krasovskii, N. N., "Approximate Computation of Optimal Control by Direct Method," *Appl. Math. and Mech.*, **24**, 390–397, 1960.

[90] Kretschmer, K. S., "Programmes in Paired Spaces," *Canadian J. Math.*, **13** (1), 221–238, 1961.

[91] Kuhn, H. W. and A. W. Tucker, "Nonlinear Programming," *Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability*, Univ. of Calif. Press, Berkeley, 1961, pp. 481–492.

[92] Kulikowski, R., "Optimizing Processes and Synthesis of Optimizing Automatic Control Systems with Nonlinear Invariable Elements," *Proc. First IFAC Congress 1960*, Butterworths, London, pp. 473–477.

[93] Lack, G. N. T., "Optimization Studies with Applications to Planning in the Electric Power Industry and Optimal Control Theory," Rept CCS-5, Institute in Engineering-Economic Systems, Stanford Univ., Aug. 1965.

[94] Lasdon, L. S., S. K. Mitter, and A. D. Waren, "The Conjugate Gradient Method for Optimal Control Problems," *IEEE Trans. on Automatic Control*, **AC-12** (2), 132–138, Apr. 1967.

[95] Lee, E. B. and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[96] Leon, A., "A Classified Bibliography on Optimization," in *Recent Advances in Optimization Techniques*, ed. by A. Lavi and T. P. Vogl, John Wiley, New York, 1965, pp. 599–645.

[97] Lorch, E. R., "Convexity and Normed Spaces," *Publications L'Institute Mathématique, Académie Serbe des Sciences*, 4, 109–112, 1952.

[98] Luenberger, D. G., "Resolution of Mass Spectrometer Data," Report SU-SEL-64-129, Stanford Electronics Lab., Stanford, Calif., Nov. 1964.

[99] Luenberger, D. G., "A Simple Polynomial Estimator," *IEEE Trans. on Automatic Control*, **AC-12**(2), 211–212, Apr. 1967.

[100] Luenberger, D. G., "Quasi-Convex Programming," *J. SIAM Appl. Math.* Sept., 1968

[101] Luisternik, L. and V. Sobolev, *Elements of Functional Analysis*, Frederick Ungar, New York, 1961.

[102] Mangasarian, O. L., "Duality in Nonlinear Programming," *Quar. Appl. Math.*, **20**, 300–302, 1962.

[103] Mann, H. B., *Analysis and Design of Experiments; Analysis of Variance and Analysis of Variance Designs*, Dover, New York, 1949.

[104] McKinsey, J. C. C., *Introduction to the Theory of Games*, McGraw-Hill, New York, 1952.

[105] Moore, E. H., *Bull. Amer. Math. Soc.*, **26**, 394–395, 1920.

[106] Moore, E. H., *General Analysis*, Part I, Memoirs, Amer. Philosophical Soc. 1, 1935.

[107] Moreau, J. J., "Fonctions convexes duales et points promimaux dans un espace hilbertien," *C.R. Acad. Sci.*, Paris, **255**, 2897–2899, 1962.

[108] Nashed, M. Z., "The Convergence of the Method of Steepest Descents for Nonlinear Equations with Variational or Quasi-Variational Operators," *J. Math. and Mech.*, **13** (5), 765–794, 1964.

[109] Natanson, I. P., *Constructive Function Theory*, Volumes I, II, III, (trans. by A. N. Obolensky and J. R. Schulenberger), Frederick Ungar, New York, 1964–1965.

[110] Neustadt, L. W., "Minimum Effort Control," *J. SIAM Control*, **1**, 16–31, 1962.

[111] Neustadt, L. W., "Optimization, A Moment Problem, and Nonlinear Programming," *J. SIAM Control*, **2**, 33–53, 1964.

[112] Neustadt, L. W., "An Abstract Variational Theory with Applications to a Broad Class of Optimization Problems. I. General Theory," *J. SIAM Control*, **4** (3), 505–527, 1966.

[113] Neustadt, L. W., "An Abstract Variational Theory with Applications to a Broad Class of Optimization Problems. II. Applications," *J. SIAM Control*, **5** (1), 90–137, 1967.

[114] Nirenberg, L., "Functional Analysis," lectures given in 1960–61, notes by Lesley Sibner, New York Univ., 1961.

[115] Ostrowski, A. M., *Solutions of Equations and Systems of Equations*, Sec. Ed., Academic Press, New York, 1966.

[116] Parzen, E., "An Approach to Time Series Analysis," *Anal. Math. Stat.*, **32** (4), 951–989, 1961.

[117] Penrose, R., "A Generalized Inverse for Matrices," *Cambridge Philosophical Soc.*, **51**, 406–413, 1955.

[118] Penrose, R., "On Best Approximate Solutions of Linear Matrix Equations," *Cambridge Philosophical Soc.*, **52**, 17–19, 1956.

[119] Pontryagin, L. S., V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes* (trans, by K. N. Trirogoff, ed. by L. W. Neustadt), Interscience, John Wiley, New York, 1962.

[120] Porter, W. A., *Modern Foundations of Systems Engineering*, Macmillan, New York, 1966.

[121] Råström, H., "Polar Reciprocity," *Seminar on Convex Sets*, Institute for Advanced Study, Princeton, N.J., 1949–1950, pp. 27–29.

[122] Rauch, H. E., F. Tung, and C. T. Striebel, "On the Maximum Likelihood Estimates for Linear Dynamic Systems," Lockheed Missiles and Space Co. Tech. Rept. No. 6-90-63-62, Palto Alto, Calif., June 1963.

[123] Riesz, F. and B. Sz-Nagy, *Functional Analysis*, Frederick Ungar, New York, 1955.

[124] Rissanen, J., "On Duality Without Convexity," *J. Math. Anal. and Appl.*, **18** (2), 269–275, May 1967.

[125] Ritter, K., "Duality for Nonlinear Programming in a Banach Space," *J. SIAM Appl. Math.*, **15** (2), 294–302, Mar. 1967.

[126] Rockafellar, R. T., "Duality Theorems for Convex Functions," *Amer. Math. Soc. Bull.*, **70**, 189–192, 1964.

[127] Rockafellar, R. T., "Extension of Fenchel's Duality Theorem for Convex Functions," *Duke Math. J.*, **33**, 81–90, 1966.

[128] Rosen, J. B., "The Gradient Projection Method for Nonlinear Programming: Part I, Linear Constraints," *J. Soc. Indust. Appl. Math.*, **8**, 181–217, 1960.

[129] Rosen, J. B., "The Gradient Projection Method for Nonlinear Programming, Part II, Nonlinear Constraints," *J. Soc. Indust. Appl. Math.*, **9**, 514–532, 1961.

[130] Rosenbloom, P. C., "The Method of Steepest Descent," *Proc. of Symposia in Appl. Math.*, **VI**, 127–176, 1956.

[131] Royden, H. L., *Real Analysis*, Macmillan, New York, 1963.

[132] Rozonoer, L. I., "Theory of Optimum Systems. I," *Automation and Remote Control*, **20** (10), 1288–1302, 1959.

[133] Rozonoer, L. I., "L. S. Pontryagin's Maximum Principle in Optimal System Theory, II," *Automation and Remote Control*, **20** (11), 1405–1421, 1959.

[134] Rozonoer, L. I., "The Maximum Principle of L. S. Pontryagin in Optimal-System Theory, III," *Automation and Remote Control*, **20** (12), 1517–1532, 1959.

[135] Russell, D. L., "The Kuhn-Tucker Conditions in Banach Space with an Application to Control Theory," *J. Math. Anal. and Appl.*, **15**, 200–212, 1966.

[136] Saaty, T. L. and J. Bram, *Nonlinear Mathematics*, McGraw-Hill, New York, 1964.

[137] Sarachik, P. E., "Functional Analysis in Automatic Control," *IEEE International Convention Record*, Part 6, pp. 96–107, 1965.

[138] Shah, B. V., R. J. Buehler, and O. Kempthorne, "Some Algorithms for Minimizing a Function of Several Variables," *J. Soc. Indust. Appl. Math.*, **12** (1), 74–91, Mar. 1964.

[139] Simmons, G. F., *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York, 1963.

[140] Sinnott, J. F. and D. G. Luenberger, "Solution of Optimal Control Problems by the Method of Conjugate Gradients," *Preprints of 1967 Joint Automatic Control Conference*, pp. 566–573.

[141] Sion, M., "On General Minimax Theorems," *Pacific J. of Math.*, **8**, 171–176, 1958.

[142] Slater, M., "Lagrange Multipliers Revisited: A Contribution to Non-Linear Programming," Cowles Commission Discussion Paper, Math. 403, Nov. 1950.

[143] Spang, H. A. III, "A Review of Minimization Techniques for Nonlinear Functions, *SIAM Rev.* **4** (4), 343, Nov. 1962.

[144] Taussky, O., "A Recurring Theorem on Determinants," *Amer. Math. Monthly*, **56**, 672–676, 1949.

[145] Taylor, A. E., *Introduction to Functional Analysis*, John Wiley, New York, 1958.

[146] Tseng, Y. Y. "Virtual Solution and General Inversions," *Uspehi Mat. Nauk. (NS)*, **11**, 213–215, 1956. (Reviewed in *Math. Rev.*, **18**, 749, 1957).

[147] Tukey, J. W., "Some Notes on the Separation of Convex Sets," *Portugaliae Math.*, **3**, 95–102, 1942.

[148] Valentine, F. A., *Convex Sets*, McGraw-Hill, New York, 1964.

[149] Varaiya, P, P., "Nonlinear Programming in Banach Space," *J. SIAM Appl. Math.*, **15** (2), 284–293, Mar. 1967.

[150] Vorobyev, Y. V., *Method of Moments in Applied Mathematics*, Gordon and Breach Science Pub., New York, 1965.

[151] Whinston, A., "Some Applications of the Conjugate Function Theory to Duality," Chapter V of *Nonlinear Programming*, ed. by J. Abadie, Interscience, New York, 1967.

[152] Wiener, N., *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, Technology Press, MIT, and John Wiley, New York, 1949.

[153] Wilde, D. J. and C. S. Beightler, *Foundations of Optimization*, Prentice-Hall, Englewood Cliffs, N.J., 1967.

[154] Wilson, R. "Computation of Optimal Controls," *J. Math. Anal. and Appl.*, **14**, 77–82, 1966.

[155] Wolfe, P., "A Duality Theorem for Nonlinear Programming," *Quar. Appl. Math.*, **19**, 239–244, 1961.

[156] Wolfe, P., "Methods of Nonlinear Programming," Chap. 6 of *Nonlinear Programming*, ed. by J. Abadie, Interscience, John Wiley, New York, 1967, pp. 97–131.

[157] Yosida, K., *Functional Analysis*, Springer-Verlag, Germany, 1965.

[158] Zoutendijk, G., *Methods of Feasible Directions*, Elsevier Pub. Co., 1960.

[159] Zoutendijk, G., "Nonlinear Programming: A Numerical Survey," *J. SIAM Control*, **4** (1), 194–210, 1966.

SYMBOL INDEX

SUBJECT INDEX